Genome Alignment, Evolution of Prokaryotic Genome Organization, and Prediction of Gene Function Using Genomic Context

Yuri I. Wolf, Igor B. Rogozin, Alexey S. Kondrashov, and Eugene V. Koonin[1]

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA

Gene order in prokaryotes is conserved to a much lesser extent than protein sequences. Only several operons, primarily those that code for physically interacting proteins, are conserved in all or most of the bacterial and archaeal genomes. Nevertheless, even the limited conservation of operon organization that is observed can provide valuable evolutionary and functional clues through multiple genome comparisons. A program for constructing gapped local alignments of conserved gene strings in two genomes was developed. The statistical significance of the local alignments was assessed using Monte Carlo simulations. Sets of local alignments were generated for all pairs of completely sequenced bacterial and archaeal genomes, and for each genome a template-anchored multiple alignment was constructed. In most pairwise genome comparisons, <10% of the genes in each genome belonged to conserved gene strings. When closely related pairs of species (i.e., two mycoplasmas) are excluded, the total coverage of genomes by conserved gene strings ranged from <5% for the cyanobacterium *Synechocystis sp* to 24% for the minimal genome of *Mycoplasma genitalium*, and 23% in *Thermotoga maritima*. The coverage of the archaeal genomes was only slightly lower than that of bacterial genomes. The majority of the conserved gene strings are known operons, with the ribosomal superoperon being the top-scoring string in most genome comparisons. However, in some of the bacterial–archaeal pairs, the superoperon is rearranged to the extent that other operons, primarily those subject to horizontal transfer, show the greatest level of conservation, such as the archaeal-type H+-ATPase operon or ABC-type transport cassettes. The level of gene order conservation among prokaryotic genomes was compared to the cooccurrence of genomes in clusters of orthologous genes (COGs) and to the conservation of protein sequences themselves. Only limited correlation was observed between these evolutionary variables. Gene order conservation shows a much lower variance than the cooccurrence of genomes in COGs, which indicates that intragenome homogenization via recombination occurs in evolution much faster than intergenome homogenization via horizontal gene transfer and lineage-specific gene loss. The potential of using template-anchored multiple-genome alignments for predicting functions of uncharacterized genes was quantitatively assessed. Functions were predicted or significantly clarified for ~90 COGs (~4% of the total of 2414 analyzed COGs). The most significant predictions were obtained for the poorly characterized archaeal genomes; these include a previously uncharacterized restriction-modification system, a nuclease-helicase combination implicated in DNA repair, and the probable archaeal counterpart of the eukaryotic exosome. Multiple genome alignments are a resource for studies on operon rearrangement and disruption, which is central to our understanding of the evolution of prokaryotic genomes. Because of the rapid evolution of the gene order, the potential of genome alignment for prediction of gene functions is limited, but nevertheless, such predictions information significantly complements the results obtained through protein sequence and structure analysis.

One of the unexpected findings of the first comparisons of complete bacterial genomes has been the near lack of gene order conservation beyond the level of operons (groups of adjacent, coexpressed and coregulated genes that encode functionally interacting proteins) even between relatively close species such as *Escherichia coli* and *Haemophilus influenzae* (Koonin et al. 1996; Tatusov et al. 1996). At an even closer evolutionary distance, such as that between *Mycoplasma genitalium* and *Mycoplasma pneumoniae*, large-scale conservation is seen, but even in this case there are several breakpoints of genome colinearity (Himmelreich et al. 1997). Subsequent comparisons of the sequenced bacterial and archaeal genomes have shown that even most of the operons are extensively rearranged during evolution. Only a few operons, typically coding for physically interacting proteins, are conserved in all or most of the genomes (Mushegian and Koonin 1996; Siefert et al. 1997; Watanabe et al. 1997; Dandekar et al. 1998; Itoh et al. 1999; Huynen and Snel 2000; 2000a). Examples include the ribosomal protein oper-

[1]Corresponding author.
E-MAIL koonin@ncbi.nlm.nih.gov; FAX (301)480-9241.
Article published on-line before print: *Genome Res.,* 10.1101/gr.161901. Article and publication are at www.genome.org/cgi/doi/10.1101/gr.161901.

11:356–372 ©2001 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/01 $5.00; www.genome.org

ons, some of which form superoperons (large arrays of genes that include several operons with a complex pattern of regulation) consisting of >20 genes, and are in general the most conserved portions of prokaryotic genomes, proton ATPases, and ABC-type membrane transport cassettes. Many individual pairs of genomes, however, share a significant number of additional operons. In part, the similarity of the gene order between prokaryotic genomes is maintained via horizontal transfer of operons, in some cases between taxonomically distant species (Itoh et al. 1999). Operon transfer is likely to be favored by selection over transfer of individual genes because, in the former case, gene coexpression and coregulation are preserved (Lawrence 1999).

Operon transfer, shuffling, disruption, and perhaps formation are major factors in the evolution of bacteria and archaea (Itoh et al. 1999; Lawrence 1997, 1999), and understanding the evolutionary dynamics of prokaryotic genomes is impossible without a detailed comparison of genome organization. On a more practical plane, conservation of gene order has been considered as one of the important predictors of gene function in prokaryotes under the general paradigm of context analysis, which is becoming increasingly popular with the growth of the collection of genome sequences (Galperin and Koonin 2000; Huynen and Snel 2000; Huynen et al. 2000a). The logic of this approach is straightforward: On one hand, by definition, genes in an operon are coexpressed and encode functionally linked proteins (Jacob et al. 1960); conversely, as indicated above, recombinational gene shuffling in prokaryotes is extensive, which results in very limited conservation of gene order between evolutionarily distant genomes. Hence, if a conserved gene string occurs in two or, better yet, three or more such genomes, there is, for all practical purposes, little doubt that the respective genes comprise an operon and are functionally linked. If one or more of these genes has no known function or has a function defined only in general terms, a prediction becomes possible. Overbeek et al. (1998, 1999) developed a simple method that enumerates conserved gene strings in pairs of genomes and provides material for such predictions. A tool for detecting conserved gene strings in pairs of genomes is also provided by the KEGG database (Kanehisa and Goto 2000).

We were interested in systematically and quantitatively exploring the conservation of gene order among the 25 currently available complete bacterial and archaeal genomes and its dependence on the evolutionary distance between these genomes, and in assessing the potential of such comparisons for identification of previously undetected operons and prediction of gene functions. These systematic genome comparisons revealed major differences between genomes in terms of their coverage with conserved operons and reinforced the notion of rapid evolution of prokaryotic gene order, which puts limits on its utility for functional prediction. Nevertheless, several previously undetected, conserved operons were identified, and functions of ~90 widespread bacterial and archaeal genes were predicted.

RESULTS AND DISCUSSION

Local Alignments of Gene Strings, Template-Anchored Multiple Genome Alignments, and Evolutionary Conservation of (Predicted) Operons

The principles of gene-by-gene genome alignment employed here were exactly the same as those of the more traditional alignment of nucleotide and protein sequences. In each case, either global or local alignments can be constructed (Needleman and Wunsch 1970; Smith and Waterman 1981; Altschul and Gish 1996). However, the global alignment approach, which is not practicable even for distantly related protein sequences, is not applicable to genome alignment at all (except, possibly, pairs of very closely related isolates of the same microbial strain) due to a large number of transpositions and inversions that occur during evolution. In practice, therefore, the task is to detect the maximal-length strings of genes with a conserved order, including possible gaps and mismatches. The construction of a gene-by-gene alignment differs from the construction of a sequence alignment in that the former requires the extra step of generating an all-against-all matrix of gene-to-gene sequence comparisons. Typically, this is done by comparing the protein sequences encoded in all genes from the analyzed genomes using, for example, the Smith-Waterman or BLAST algorithms. Here we compared protein sequences encoded in all completely sequenced prokaryotic genomes using the BLASTP program and converted the resulting scores into density values to make the scores independent from protein lengths (see Methods). Scores for gene pairs in a genome-to-genome comparison were then extracted from the sequence comparison matrix by one of two approaches: (1) using either *0* and *1* values only, with a value of *1* assigned to all bidirectional genome-to-genome best hits and a value of *0* assigned to all other gene pairs; or (2) using score density values for each protein pair. This is generally analogous to the use of a simple identity matrix versus a residue substitution matrix such as PAM or BLOSUM in protein sequence alignment (Henikoff and Henikoff 2000), but, unlike the sequence case, the genome alignments produced using these two approaches differ not only in extent, but in substance. The alignment constructed using the bidirectional best hits is likely to consist mostly, if not exclusively, of pairs of orthologous genes (Fitch 1970;

Tatusov et al. 1997) and, by extension, of orthologous gene strings (potential operons). In contrast, a genome alignment constructed with the use of similarity scores for gene pairs will include paralogous operons, the existence of which is well documented, the ABC-type transport cassettes being perhaps the best example (Tomii and Kanehisa 1998). The two types of alignment appear to be optimal for different purposes and were used accordingly here. An alignment that includes only orthologs is most appropriate when evolutionary conservation of gene order is being explored; in contrast, when the complete coverage of a genome with conserved gene strings is examined and functional predictions are attempted, it is advantageous to include paralogs.

As with sequence alignment, gene-by-gene genome alignment requires that gap and mismatch penalties are introduced, for which no solid theoretical basis is available, and therefore a degree of arbitrariness is inevitable. Typically, protein sequence alignment methods use a relatively high gap-opening penalty and a significantly lower gap-extension penalty; the reasoning behind this is that inserts/deletions affect protein structure and therefore are tolerated only to a limited extent, but once such a mutation is fixed, further changes of the insert length are of lesser consequence, with even very long inserts present in many proteins (Vingron and Waterman 1994; Altschul 1998). These considerations do not apply to genome evolution where the probability of functional association between (formerly) adjacent genes being maintained is expected to decrease rapidly with the increase in the number of inserted genes. Therefore, in the present analysis, we used a gap-opening penalty of zero and a linear function for the gap/mismatch extension penalty (a mismatch, i.e., a pair of genes whose products did not show significant sequence similarity to one another, was treated as equivalent to a gap [Altschul 1998]).

The program `Lamarck` (see Methods) was used to produce local gene-by-gene alignments for all pairs of bacterial and archaeal genomes using each of the above approaches. The gap/mismatch penalties were selected empirically so that the known large gene clusters that are known to be subject to rearrangement, primarily the ribosomal superoperon, were detected in their entirety. The score cut-off for the detection of conserved gene strings was naturally set at two for the *0–1* scoring scheme so that each pair of apparent orthologs in a row was reported, and the cut-off for the information-density-based scheme was similarly adjusted to include pairs of homologs with significant sequence similarity. The statistical significance (expressed as the random expectation [*E*] value) of the detected strings of homologous genes was estimated using Monte Carlo simulations.

The inherent limitation of this method of gene order analysis is that completely reshuffled, but nevertheless conserved operons will not be identified. For a conserved gene string to be detected, at least one pair of homologous with the same gene order (but possibly with an inserted gene[s]) is required. The method of Overbeek and co-workers for detection of clusters of potential functionally linked genes is, in this respect, more general because conserved clusters are identified on the basis of genes belonging to the same run, regardless of the exact gene order (Overbeek et al. 1999). The latter method, however, involves other, more or less arbitrary assumptions, namely that only genes transcribed in the same direction and separated by a distance not exceeding a certain maximal number of base pairs (300 in this particular study) are considered (see below).

The data in Table 1 show the dependence of the number of detected conserved gene strings and the number of genes in them on the *E*-value cut-off for the two scoring schemes. Generally, the presence of a pair of adjacent homologous genes in two genomes is not statistically significant because the number of detected pairs is much greater at $E < 0.1$ than at $E < 0.01$ (Fig. 1). In other words, many such pairs occur simply by chance, whereas others are functionally and evolutionarily relevant; the two situations can be distinguished by analyzing multiple genomes and/or by examining gene pairs case-by-case. Indeed, if the gene orders in different genomes are considered independent, which for evolutionarily distant genomes might be a reasonable approximation, the probabilities of the occurrence of gene pairs should be multiplied, and accordingly, the presence of a spurious pair of adjacent homologs in four or more genomes is extremely unlikely. The characteristic length of a nonrandom, statistically significant ($E < 0.01$) conserved gene string in most genome pairs is commensurate with the characteristic operon length, namely 3–4 genes (Fig. 1B). The genome alignment method employed here does not require that genes in a string are transcribed in the same direction or down-weigh gene pair that are transcribed in different directions. Nevertheless, nearly all conserved gene pairs detected in genomes other than closely related ones, particularly at $E < .01$, are unidirectional ones, which is compatible with the notion that such gene pairs belong to conserved operons (Fig. 2).

Most of the pairwise genome comparisons reveal only one long conserved string, which, not unexpectedly, consists of varying substrings of the ribosomal superoperon (Fig. 1, Table 2). In several bacterial–archaeal pairings, however, the super-operon is disrupted to the extent that other operons, such as the archaeal-type ATPase operon, appear as the highest-scoring local alignment, which usually, but not always (because of gaps and mismatches), is also the longest

Table 1. Conserved Gene Strings in Pairwise Genome Comparisons

		# conserved strings / # aligned genes / % in genome 1 / % in genome 2			
		Alignment method			
		All hits (information density)		Bidirectional best hits (orthologs)	
Genome 1	Genome 2	E < 0.1	E < 0.01	E < 0.1	E < 0.01
Ctra	Cpneu	33/745/83%/71%	31/738/83%/70%	35/757/85%/72%	34/753/84%/72%
Tpal	Bbur	27/161/16%/19%	27/161/16%/19%	31/176/17%/21%	31/176/17%/21%
Ecoli	Hinf	138/566/13%/33%	80/411/10%/24%	105/482/11%/28%	105/482/11%/28%
Ecoli	Bsub	89/322/8%/8%	36/182/4%/4%	34/168/4%/4%	34/168/4%/4%
Bsub	Synecho	16/74/2%/2%	9/58/1%/2%	29/94/2%/3%	9/50/1%/2%
Synecho	Aquae	20/67/2%/4%	14/53/2%/3%	9/40/1%/3%	9/40/1%/3%
Aquae	Tmar	13/56/4%/3%	7/41/3%/2%	12/58/4%/3%	12/58/4%/3%
Tmar	Mtub	30/129/7%/3%	14/85/5%/2%	46/150/8%/4%	17/92/5%/2%
Tmar	Drad	37/148/8%/5%	27/125/7%/4%	10/58/3%/2%	10/58/3%/2%
Ecoli	Mjan	10/30/1%/2%	5/18/0%/1%	12/33/1%/2%	5/19/0%/1%
Ecoli	Aero	54/191/4%/10%	36/145/3%/8%	25/56/1%/3%	5/16/0%/1%
Tmar	Mjan	7/27/1%/2%	4/19/1%/1%	26/64/3%/4%	7/30/2%/2%
Tmar	Aero	58/220/12%/12%	46/196/11%/11%	31/85/5%/5%	12/47/3%/3%
Mjan	Aero	18/68/4%/4%	17/66/4%/4%	11/55/3%/3%	11/55/3%/3%
Mjan	Mthe	41/184/11%/10%	32/163/10%/9%	28/148/9%/8%	28/148/9%/8%
Pyro	Aero	29/122/6%/7%	17/90/4%/5%	11/60/3%/3%	11/60/3%/3%

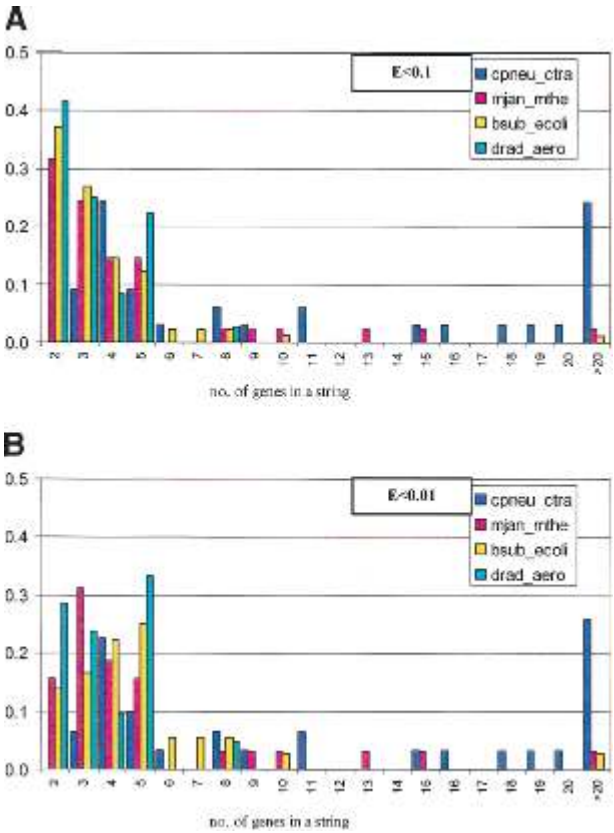


Figure 1 Length distribution of conserved gene strings in pairs of prokaryotic genomes. The distributions are for information-density-based alignments; the respective *E*-value cut-offs are indicated in panels *A* and *B*. The vertical axis shows the fraction of aligned gene strings for a given genome pair.

conserved gene string (Table 2). Most of the alignments between distant genomes did not include any strings spanning two or more functionally unrelated operons, which would be interpreted as preservation of the ancestral gene order.

Typically, pairwise genome alignments included a small fraction of genes in each of the compared genomes, on most occasions <10% (Table 1). Only for pairs of closely related genomes, such as two species of *Chlamydia* or two species of *Mycoplasma*, the fraction of the genes included in alignments was significantly greater (Table 1). In contrast, some of the pairwise genome alignments, for example, those between some of the bacteria and the archaeon *Methanococcus jannaschii*, showed an extremely low overlap, with <1% of the genes involved and <10 conserved gene strings (Table 1). It is worth noting that, in some of the com-

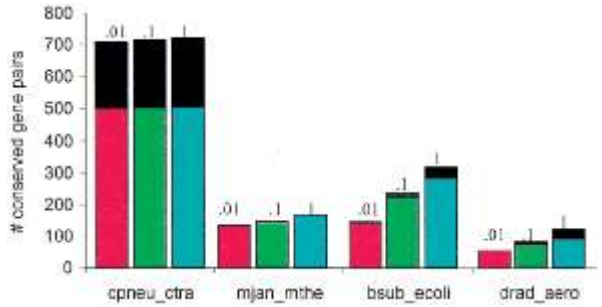


Figure 2 The prevalence of gene pairs transcribed in the same direction among conserved gene strings. For each pair of genomes, the number of unidirectional gene pairs (colored) and gene pairs transcribed in the opposite direction (black) is shown for three *E*-value cut-offs (indicated on top of each bar).

Table 2. The Most Conserved Gene Strings in Pairwise Genome Comparisons

	Best local alignment, length, number of gaps/mismatches[a]					
	E. coli	*B. subtilis*	*D. radiodurans*	*Synechocystis sp.*	*A. aeolicus*	*A. pernix*
B. subtilis	r-superoperon, 30, 2					
D. radiodurans	r-superoperon, 14, 0	r-superoperon, 19, 1				
Synechocystis sp.	r-superoperon, 27, 4	r-superoperon, 35, 4	r-superoperon, 13, 0			
A. aeolicus	r-superoperon, 9, 0	r-superoperon, 10, 0	r-superoperon, 10, 0	r-superoperon, 10, 0		
A. pernix	Phosphate transport operon, 4, 0	r-superoperon, 12, 2	Phosphate transport operon, 5, 0	r-superoperon, 3, 0	r-superoperon, 5, 0	
M. jannaschii	Phosphate transport operon, 4, 0	r-superoperon, 5, 0	Archaeal-type H+-ATPase operon, 8, 0	r-superoperon, 10, 2	r-superoperon, 5, 0	r-superoperon, 16, 1

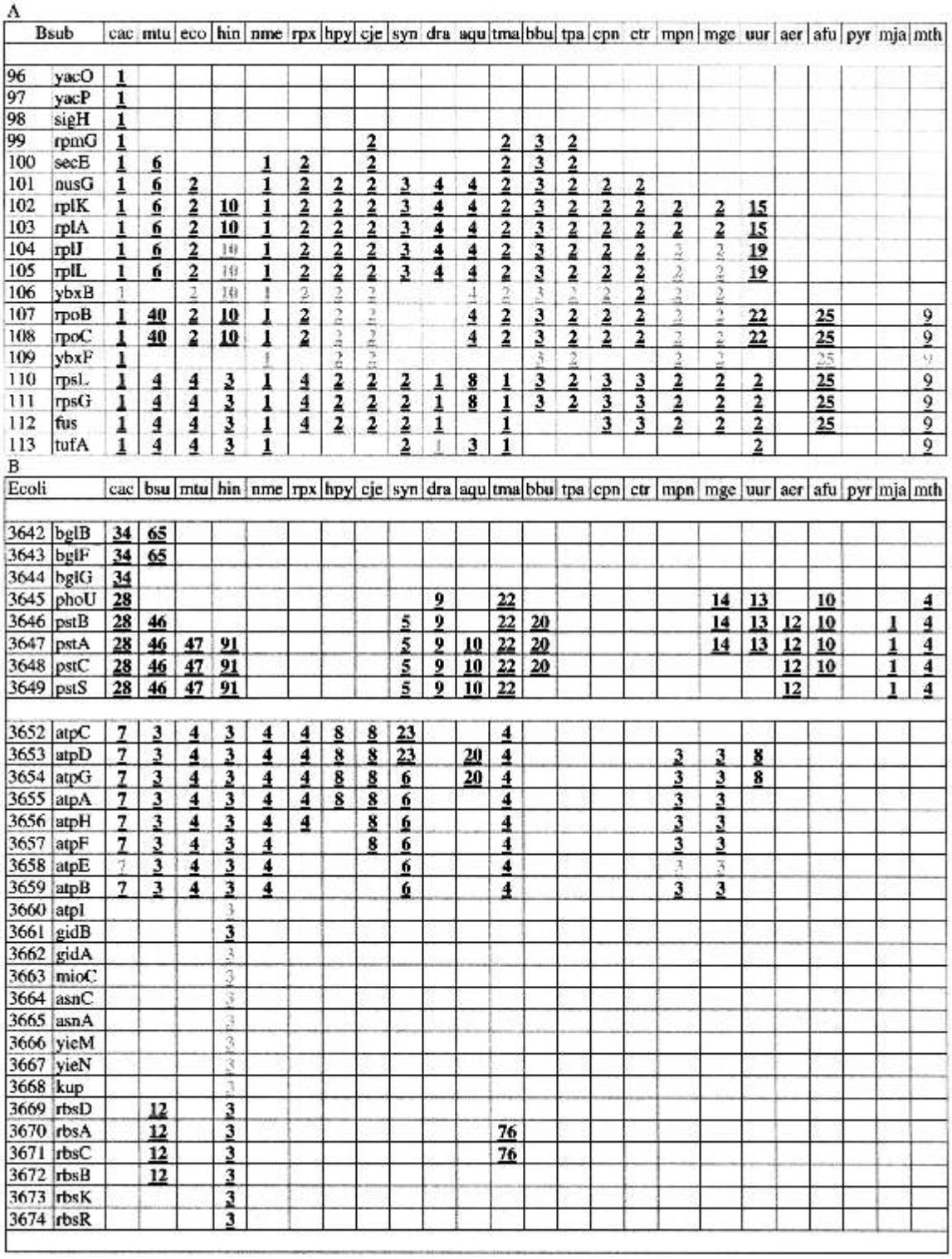
[a]The data are for the information-density-based alignments.

parisons between distantly related genomes, the bidirectional best hit method of alignment construction revealed additional conserved gene strings (e.g., the *Bacillus subtilis–Synechocystis* and *Theomotoga maritima– M. jannaschii* comparisons in Table 1). This occurred because under the information-density-based method, strings of orthologous genes with low sequence similarity may not be detected or may fail to pass the *E*-value cut-off.

Taken together, these findings are in accord with the notion that gene (operon) order is poorly conserved among bacteria and archaea. Moreover, they strongly suggest that, at large evolutionary distances, the genomes are (nearly) homogenized with respect to the operon arrangement, with (virtually) no memory of the ancestral genome organization. A corollary to this conclusion is that statistically significant conserved gene strings can be confidently predicted to form operons. More specifically, this notion is supported by at least five lines of argument: (1) The likelihood that such conserved gene strings are observed by chance is low; (2) there are few of such conserved strings, and all of those that include functionally characterized genes correspond to known or predicted (on the basis of the obviously linked functions of the constituent genes) operons (see the complete results at ftp://ncbi.nlm.nih.gov/pub/koonin/genome_align); (3) typical conserved gene strings include 3–4 genes which is also the characteristic size of operons; (4) conserved gene strings that include genes from adjacent, independent operons are extremely rare; (5) nearly all conserved gene strings consist of genes that are transcribed in the same direction.

Given the limited conservation of the gene order detected in pairwise genome alignments, we combined them to produce template-anchored multiple alignments for each of the bacterial and archaeal genomes. A template-anchored alignment shows the total coverage of the given genome (the template) with conserved gene strings that comprise each of the pairwise alignments (Fig. 3). Even in these multiple alignments, the total coverage of genomes with conserved gene strings (known and predicted operons) was generally low, but showed a striking range from <5% for the cyanobacterium *Synechocystis sp*. to 24% in *M. genitalium* (the bacterium with the smallest genome) and 23% for the hyperthermophilic bacterium *Thermotoga maritima* (Fig. 4). The paucity of conserved operons in *Synechocystis* has been noticed previously (Itoh et al. 1999). An attractive recent hypothesis postulates origin of operons encoding macromolecular complexes in ancestral hyperthermophilic prokaryotes under the selective pressure to channel thermolabile substrates (Glansdorff 1999). Under this scenario, it could be expected that genomes of hyperthermophiles would show a greater density of operons than those of mesophiles. Clearly, however, there is no such consistent trend, at least with respect to operons that are conserved in evolution; indeed, whereas *T. maritima* is among the genomes most densely covered with conserved operons, the other bacterial hyperthermophile, *Aquifex aeolicus*, is at the lower end of the spectrum (Fig. 4). Archaea also show relatively low coverage in spite of the availability of six archaeal genomes (Fig. 4). Thus the analysis of gene order conservation fails to yield evidence in support of the thermophilic hypothesis of operon origin.

The number of potential operons in a prokaryotic genome may be approximated by a number of gene strings that are transcribed in the same direction (such groups of genes have been aptly dubbed 'directons',



A

	Bsub	cac	mtu	eco	hin	nme	rpx	hpy	cje	syn	dra	aqu	tma	bbu	tpa	cpn	ctr	mpn	mge	uur	aer	afu	pyr	mja	mth
96	yacO	1																							
97	yacP	1																							
98	sigH	1																							
99	rpmG	1							2				2	3	2										
100	secE	1	6			1	2		2				2	3	2										
101	nusG	1	6	2		1	2	2	2	3	4	4	2	3	2	2	2								
102	rplK	1	6	2	10	1	2	2	2	3	4	4	2	3	2	2	2	2	2	15					
103	rplA	1	6	2	10	1	2	2	2	3	4	4	2	3	2	2	2	2	2	15					
104	rplJ	1	6	2	10	1	2	2	2	3	4	4	2	3	2	2	2	2	2	19					
105	rplL	1	6	2	10	1	2	2	2	3	4	4	2	3	2	2	2	2	2	19					
106	ybxB	1		2	10	1	2	2	2			4	2	3	2	2	2	2	2						
107	rpoB	1	40	2	10	1	2	2	2			4	2	3	2	2	2	2	2	22		25			9
108	rpoC	1	40	2	10	1	2	2	2			4	2	3	2	2	2	2	2	22		25			9
109	ybxF	1				1		2	2					3	2			2	2			25			9
110	rpsL	1	4	4	3	1	4	2	2	2	1	8	1	3	2	3	3	2	2	2		25			9
111	rpsG	1	4	4	3	1	4	2	2	2	1	8	1	3	2	3	3	2	2	2		25			9
112	fus	1	4	4	3	1	4	2	2	2	1		1			3	3	2	2	2		25			9
113	tufA	1	4	4	3	1				2	1	3	1							2					9

B

	Ecoli	cac	bsu	mtu	hin	nme	rpx	hpy	cje	syn	dra	aqu	tma	bbu	tpa	cpn	ctr	mpn	mge	uur	aer	afu	pyr	mja	mth
3642	bglB	34	65																						
3643	bglF	34	65																						
3644	bglG	34																							
3645	phoU	28									9		22						14	13		10			4
3646	pstB	28	46							5	9		22	20					14	13	12	10		1	4
3647	pstA	28	46	47	91					5	9	10	22	20					14	13	12	10		1	4
3648	pstC	28	46	47	91					5	9	10	22	20							12	10		1	4
3649	pstS	28	46	47	91					5	9	10	22								12			1	4
3652	atpC	7	3	4	3	4	4	8	8	23			4												
3653	atpD	7	3	4	3	4	4	8	8	23		20	4					3	3	8					
3654	atpG	7	3	4	3	4	4	8	8	6		20	4					3	3	8					
3655	atpA	7	3	4	3	4	4	8	8	6			4					3	3						
3656	atpH	7	3	4	3	4	4		8	6			4					3	3						
3657	atpF	7	3	4	3	4			8	6			4					3	3						
3658	atpE	7	3	4	3	4				6			4					3	3						
3659	atpB	7	3	4	3	4				6			4					3	3						
3660	atpI				3																				
3661	gidB				3																				
3662	gidA				3																				
3663	mioC				3																				
3664	asnC				3																				
3665	asnA				3																				
3666	yieM				3																				
3667	yieN				3																				
3668	kup				3																				
3669	rbsD		12		3																				
3670	rbsA		12		3								76												
3671	rbsC		12		3								76												
3672	rbsB		12		3																				
3673	rbsK				3																				
3674	rbsR				3																				

Figure 3 Segments of template-anchored, gene-by-gene genome alignments. (*A*) Template *Bacillus subtilis*. A section of the ribosomal superoperon is shown. (*B*) Template *Escherichia coli*. Two distinct gene strings separated by blank lines are shown. The top string includes the β-glucosidase operon and the phosphate transport operon. The bottom string includes the H^+-ATPase operon, three genes implicated in cell division and potentially forming an operon (*gidA*, *gidB*, *mioC*), the asparagine synthetase operon, the predicted Mg-chelatase operon (*yieN-yieM*; see Table 3), and the ribose transport operon. The first column shows the number of the respective gene in the genome and the second column shows the gene name. The rest of the columns show the rank of the respective gene string in the corresponding pairwise genome comparison (information density-based alignments; *E*-value < 0.1). The gray numbers indicate positions with gaps or mismatches in the gene strings.

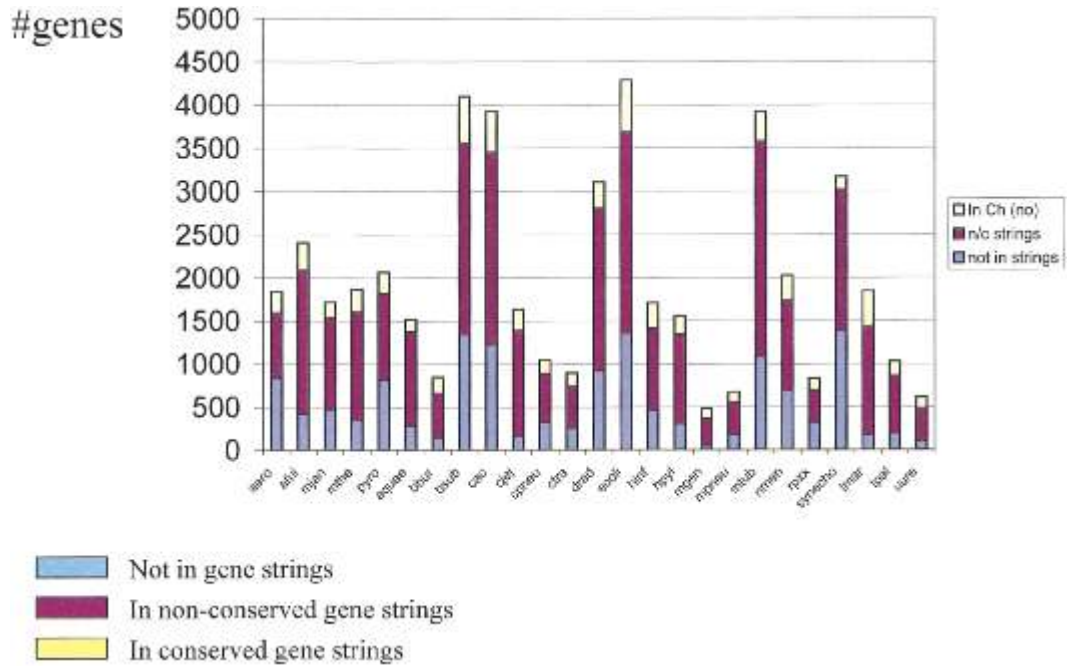


Figure 4 Coverage of prokaryotic genomes with nonconserved and conserved gene strings. Conserved gene strings were from information-density-based alignments with $E < 0.1$. The union of nonconserved and conserved gene strings was considered to comprise the set of potential operons (see text). Pairs of closely related genomes, namely *C. trachomatis/C. pneumoniae*, *M. genitalium/M. pneumoniae*, *B. subtilis/C. acetobutilicum*, and *M. jannaschii/M. thermoautotrophicum*, were disregarded for these calculations.

Salgado et al. 2000) and are separated by (relatively) short untranslated spacers. Under a liberal spacer length cut-off of 100 bp, from 54% (the crenarchaeon *Aeropyrum pernix*) to 90% (*T. maritima*) of the genes in each genome belong to tightly spaced directons, or potential operons. A recent estimate based on a detailed analysis of the distributions of spacer lengths between genes within known operons and those between transcriptional units suggested that *E. coli* could possess ~700 operons (Salgado et al. 2000), which, with the average number of 3–4 genes per operon, amounts to 2000–2500 (~55% genes). This is somewhat lower than, but not incompatible with the above estimates based on the number of directons that therefore may be used as a reasonable upper bound on the number of operons in prokaryotic genomes. Obviously, only a relatively small fraction of these potential operons show evolutionary conservation within the presently available sample of genomes (Fig. 4). Thus, either the majority of closely spaced directons are, in fact, not operons or, more likely, most of the operons are relatively unstable in evolution and are conserved only in the genomes of closely related species, or due to horizontal transfer. Retracing and simulating the effect of the accumulation of complete prokaryotic genome sequences on the coverage of genomes with conserved gene strings (predicted operons) seems to corroborate the latter view. On average, each added genome makes an incremental, unique contribution to the coverage (Fig. 5). Clearly, however, coverage grows slower than linearly, and although the available data did not allow a reliable extrapolation, it appears that, unless series of closely related genomes are sequenced, many more distantly related ones needed to approach complete coverage of all potential operons.

The coverage of a genome with conserved gene strings (predicted operons) is one characteristic of the conservation of genome organization, and the depth of coverage (in other words, the height of the stack of aligned genomes in a template-anchored alignment, as shown in Fig. 2) is another. There is a clear correlation between the two values, with *Synechocystis* showing the lowest and *M. genitalium* the highest values of both parameters; *T. maritima* stands out, with high coverage and low depth (Fig. 6A). When the alignment depth is measured for those genes that belong to conserved gene strings only, an inverse correlation is seen in that the genomes covered sparsely with conserved operons (such as *Synechocystis*) show high values of relative depth, and vice versa (Fig. 5B). In other words, those genomes that are covered sparsely with conserved gene strings encompass primarily the most common operons. This plot, however, reveals an additional anomaly in a subset of parasitic bacteria with small genomes, namely spirochetes and particularly *Rickettsia*, which show relatively broad and also deep coverage of the genome

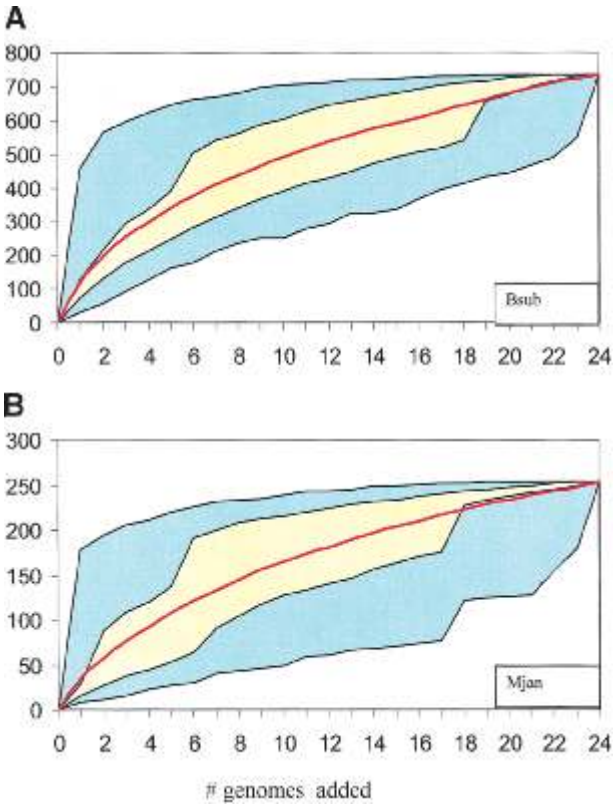


Figure 5 Contribution of accumulating genome sequences to the coverage of genomes with conserved gene strings. For a given template genome, the other 24 genomes were added one-by-one in a random order, and the number of genes from the template genome covered with conserved gene strings (information-density-based alignments, $E < 0.1$) was recorded at each step. The blue area shows the range between the minimal and the maximal coverage obtained in 100 random replications of the procedure. The yellow area shows the range between the 25% and 75% quantiles for each step. The red line shows the average alignment coverage in the 100 replications.

with conserved gene strings (Fig. 5B). Notably, the coverage of the archaeal genomes is within the range characteristic of bacteria, albeit close to its lower end (Fig. 5A,B).

The present, systematic comparative assessment of gene order conservation in bacteria and archaea indicates that: (1) at large evolutionary distances, genome rearrangement has reached — or at least is approaching — saturation, and what conservation is observed is dictated by functional constraints on operon structure; and (2) prokaryotic genomes differ dramatically in the level of (predicted) operon conservation, with the genome coverage by conserved gene strings varying from <5% to ~25%; (3) the proportion of conserved gene strings in archaeal genomes is comparable to that in bacteria; however, only a small number of (predicted) operons are shared by archaea and bacteria.

Conservation of Gene Order versus Other Measures of Genome Evolution

Conservation of gene order is one of the genome-scale evolutionary parameters that have become measurable with the availability of multiple complete genome sequences. We examined the relationship between the conservation of (predicted) operon structure and the conservation of gene repertoires and protein sequences themselves. The Clusters of Orthologous Groups of proteins (COGs) were used as the source of data on the conservation of gene repertoires (Tatusov et al. 1997, 2000). Distinct relationships were revealed between gene repertoire conservation and the conservation of gene order for the archaeal–bacterial, bacterial–bacterial and archaeal–archaeal genome comparisons. Figure 7A shows a plot of the co-occurrence coefficients between genomes in the COG set versus those in the set of conserved gene strings (see Methods). The archaeal–bacterial comparison data appear to scatter randomly in terms of gene string conservation, but are clearly separated into three distinct subsets along the COG co-occurrence axis (Fig. 7A). The subset with the lowest co-occurrence in the COGs includes the pairs of archaeal genomes and those of highly degraded bacterial parasites, namely *Mycoplasma*, *Chlamydia*, *Rickettsia* and Spirochetes; the subset with the highest co-occurrence coefficients surprisingly consists of the comparisons between the Crenarchaeon *A. pernix* and all bacterial genomes; the largest subset, which includes the rest of the interkingdom pairs, significantly overlaps with the bacterial–bacterial comparisons (Fig. 7A). The archaeal–archaeal genome pairs form a distinct set in the upper right part of the plot, with high

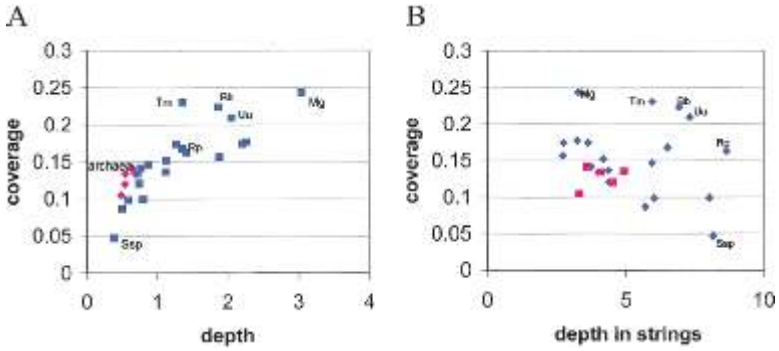


Figure 6 Conserved gene strings in prokaryotic genomes: coverage versus depth. Information-density-based alignments ($E < 0.1$) were used to calculate coverage and alignment depth. The coverage is expressed as the fraction of genes in a genome that are included in conserved gene strings. The depth is expressed as the average number of genes in a column of the template-anchored alignment for the respective genome (*A*) or the average number of genes in a column that belongs to a conserved gene string (i.e., contains at least one gene) (*B*). Abbreviations: Bb, *Borrelia burgdorferi*, Mg, *Mycoplasma genitalium*; Ssp, *Synechocystis sp.*, Tm, *Thermotoga maritima*, Uu, *Ureaplasma urealyticum*.

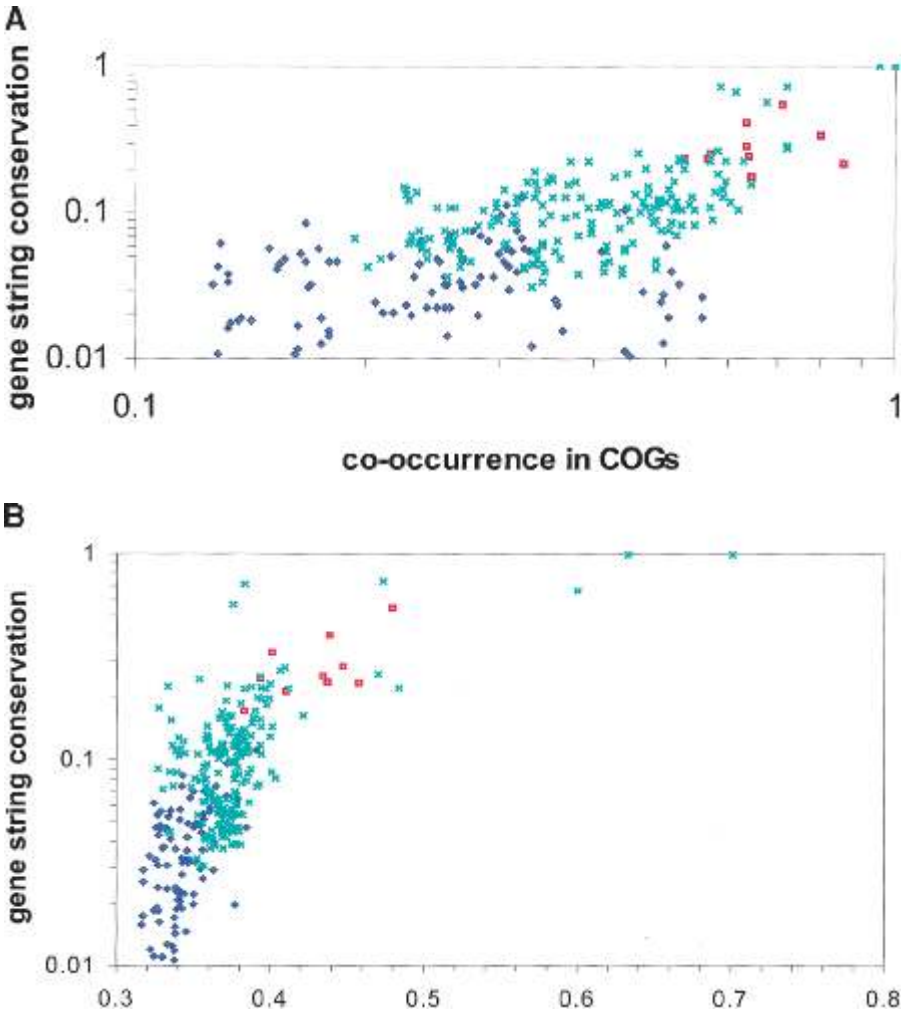


Figure 7 Measures of genome evolution: operon conservation versus conservation of gene repertoire and protein sequences. Alignments including apparent orthologs (bidirectional best hits) were used in both plots ($E < 0.1$). (*A*) Gene string conservation versus co-occurrence in clusters of orthologous genes (COGs). The co-occurrence coefficients for genome alignments and for the COGs were calculated as indicated in Methods. (Dark-blue diamonds) Archaeal-bacterial comparisons, (light-blue crosses) bacterial–bacterial comparisons, (red squares) archaeal–archaeal comparisons. (*B*) Gene string conservation versus sequence identity level between probable orthologs. Horizontal axis: median fraction of identical amino acid residues among probable orthologs. The other designations are as in *A*.

co-occurrence coefficients in both the COGs and conserved gene strings (Fig. 7A). The bacterial–bacterial genome comparisons span a wide range of values of both measures, and a correlation between the conservation of gene repertoire and that of the gene order was discernible (Fig. 7B).

The median of the similarity score (in the simplest case, percent identity) distribution among orthologs appears to be a useful measure of genome evolution that can be employed to construct phylogenetic trees (Grishin et al. 2000). When the median percent identity values were plotted against the co-occurrence co efficient in conserved gene strings, a positive correlation was observed, with a clear separation of archaeal–bacterial, bacterial–bacterial and archaeal–archaeal genome pairs (Fig. 7B). The general inverse correlation between phylogenetic distance and gene order conservation has been discussed previously (Mushegian and Koonin 1996; Siefert et al. 1997; Watanabe et al. 1997; Dandekar et al. 1998; Itoh et al. 1999).

To summarize, conservation of predicted operons (gene strings) in prokaryotic genomes showed a degree of correlation with the other evolutionary parameters we examined, namely the conservation of gene repertoires and the protein sequences themselves. However, compared to these measures, operon conservation appeared to show less correlation with phylogenetic relationships or the known features of organisms' lifestyles. It is, for example, unclear why the cyanobacterium *Synechocystis sp*. possesses so few conserved operons, and the hyperthermophilic bacterium *T. maritima* has so many, whereas the other bacterial hyperthermophile, *Aquifex*, is among the genomes with the lowest level of operon conservation. If the functional cognates of these and other peculiarities of the patterns of gene order conservation could be identified, this might shed an unexpected light on bacterial and archaeal biology.

Prediction of Gene Functions Using Information Extracted from the Gene Order Conservation

Conservation of operon organization is one of the principal types of context information, which appears to become increasingly important with the growth of the collection of completely sequenced genomes (Galperin and Koonin 2000; Huynen and Snel 2000a; Huynen et al. 2000). The results of the systematic analysis of the conservation of gene order presented here support the notion that most, if not all, gene strings that are conserved in taxonomically distant prokaryotes are indeed operons, rather than remnants of the ancestral gene order. Accordingly, the genes

found in such strings, particularly in multiple genomes, can be legitimately assumed to be functionally linked, and the information on gene clustering can be used for functional prediction. However, the very same evolutionary force that seems to ensure the robustness of predictions based on gene order conservation — rampant recombination that leaves intact only gene strings (operons) stabilized through selection — limits the potential of this source of information in terms of functional prediction.

We were interested in benchmarking the approach to gene function prediction based on gene string conservation by examining, in detail, the template-anchored genome alignments. What we sought to identify were not all predictions that could be made on the basis of gene order conservation, but those that appeared to be unique to this approach, i.e., could not be attained (at least not with comparable confidence) by sequence analysis, primarily as encapsulated in the COG database. Furthermore, we took a conservative approach by considering only those predictions that stemmed from operon conservation in at least three genomes, or statistically significant local alignments of two genomes. Only those predictions were considered that appeared to significantly contribute to our understanding of the probable function(s) of the protein in question; predictions of previously undetected functional links between proteins with well-characterized functions were not included unless deemed critical for understanding the central function of a protein. Because of all these restrictions, the set of predictions obtained here should be considered a low bound on the predictive potential of gene order comparisons. As a feedback, the proteins whose functions were predicted or clarified on the basis of operon conservation analysis were subjected to additional, detailed sequence analysis using, primarily, the PSI-BLAST program, to further enhance the predictions.

We found that functional assignments for completely uncharacterized proteins or a major clarification of the probable function were possible for ~90 COGs, or ~4% of the 2422 analyzed COGs (Table 3). Even with all the qualifications discussed above, this number should be considered an approximation because the decision on what constitutes a major prediction inevitably includes a subjective element. This uncertainty notwithstanding, the yield of unique functional predictions from operon conservation analysis was relatively modest. However, this should not detract from the apparent importance of many of these predictions. Examples include the prediction of several previously undetected components of the translation machinery, such as probable translation factors, ribosome-associated proteins and RNA modification enzymes, and components of the DNA repair machinery (Table 3). The prediction rate was nonuniform across the range of the available genomes, with more predictions attainable for the poorly characterized archaeal genomes than for bacterial genomes. The distribution of the predictions across the range of cellular functions was also nonuniform. In particular, a disproportionately large number of predictions were for translation-related and replication/repair-related functions, apparently because operons coding for components of these systems show greater evolutionary conservation than those coding for other types of cellular functions.

The predictions stemming from gene order conservation differ in character from those made using sequence-structure comparison. The latter tend to predict the biochemical activity of a protein, such as ATPase or nuclease, and even the details of the active center, but is often less informative in terms of pinpointing the protein's actual cellular role. In contrast, analysis of the gene order helps placing an uncharacterized protein in the context of a cellular functional system, such as translation or replication, but may not decipher the exact biochemistry. Combined, the analyses of sequences and structures and of gene order are capable of producing detailed functional predictions. This can be exemplified by four uncharacterized, highly conserved proteins that are predicted to possess GTPase activity on the basis of sequence analysis and are confidently assigned to the translation system on the basis of the genome context (Table 3). One of these GTPases (COG0012) is ubiquitous in all cellular life forms, and two others are present in all or nearly all bacteria (Table 3), which are the phyletic patterns characteristic of the components of the translation machinery. Furthermore, two of these predicted new translation factors (COG0012 and COG0536) additionally contain the tRNA-binding TGS (Threonyl-tRNA synthetase, GTPases, SpoT) domain (Wolf et al. 1999), which further supports the translation connection. These predictions emphasize that fundamental aspects of central cellular functions such as translation still remain to be uncovered, and gene order conservation, its limitations notwithstanding, can be one of the important sources of information for the identification of candidate proteins.

As befits an approach based on the conservation of gene strings, genome alignment analysis resulted in the prediction of several operons that encode for previously undetected functional complexes. These include two distinct restriction-modification systems, two unique helicase-nuclease combinations that probably comprise novel repair complexes, an operon that, on the basis of the predicted activities of the respective gene products, is predicted to encode a previously undetected molecular chaperone complex associated with the translation machinery (see also below), and three predicted operons that encode the probable archaeal counterpart of the eukaryotic exosome (Table 3;

Table 3. Prediction of Gene Function on the Basis of the Gene Order Conservation[a]

COG	Current functional assignment[b]	Phyletic pattern (species represented in the COG)	New functional prediction[c]	Support from (predicted) operon arrangement[d]
2501	Uncharacterized conserved protein	Ecoli, Nmen, Drad, Bsub, Uure	Small nucleic acid-binding protein, function in replication/repair	DnaA, dnaN, recF in Bsub, Cace, Mtub, Uure
0718	Uncharacterized conserved protein	All bacteria except for Aquae, Tpal	Function in replication/repair	DnaX, RecR in Bsub, Cace, Ecoli, Mtub, Uure, Drad
1273	Uncharacterized conserved protein	Aful, Bsub, Mtub	An enzyme (possibly a nuclease) functionally associated with ATP-dependent DNA ligase and eukaryotic-type DNA primase	Primase-ligase (fusion protein) in Bsub Ligase in Aful
1623	Nucleic acid-binding protein (contains the helix-hairpin-helix domain)	Tmar, Bsub, Cace, Mtub	Enzyme involved in repair, possibly a novel DNase	Sms in Tmar, Bsub, Cace, Mtu
0452	Flavoprotein involved in panthothenate metabolism	All species except Uure, Mgen, Mpne, Tpal, Ctra, Cpne, Rpro	Function in DNA replication/repair	RadC, dut in Ecoli, Hinf PriA in Cace Mutant is deficient in DNA synthesis (Spitzer and Weiss 1985)
2137	Uncharacterized conserved protein	Tmar, Drad, Ecoli, Hinf, Nmen, Bsub, Cace, Mtub, Tpal	Function in repair	RecA in Eco, Tmar Sms in Tpal
1418	HD superfamily hydrolase	All bacteria and archaea except Aper, Synecho, Hinf, Mtub, Ctra, Cpne, Rpro	Nuclease, function in recombination/repair	RecA in Bsub, Tmar
0424	Nucleotide-binding protein	All bacteria except Hinf, Uure, Mgen, Mpne, Bbur, Tpal; Phor	NTPase, function in repair	RadC in Bsub, Cace, Tmar
1546	Uncharacterized conserved protein	All bacteria except Hinf, Rpro, Bbur, Tpal, Ctra, Cpne	Function in repair/recombination	RecA in Eco, Vcho, Bsub, Uure; Operon structure in *Streptococcus pneumoniae* demonstrated (Martin et al. 1995); effect of mutation on DNA repair shown in Drad (Narumi et al. 1999)
1630	Uncharacterized conserved protein	All archaea	Function in repair, possibly a nuclease	SbcC, SbcD (ATPase and nuclease involved in repair) in Aful, Phor, Aper
0061	ATPase or kinase distantly related to phosphofructokinase	All archaea and bacteria except Drad, Ctra, Cpne	ATPase (molecular chaperone?) involved in repair	RecN, grpE in Ecoli, Vcho, Hinf, Mtub, Xfas, Hpyl, Cjej, Tpal
1518	**Uncharacterized conserved protein**	**Aful, Mjan, Mthe, Phor, Aper, Aqua, Tmar, Ecoli, Mtub, Cjej**	**Component of a novel repair system**	**Predicted operon (partially) conserved in Aqua, Tmar, Aful, Phor, Mjan, Mthe**
1468	**Metal-binding, possibly nucleic acid-binding protein**	**Aful, Mjan, Mthe, Phor, Aper, Aqua, Tmar**	**RecB-family exonuclease, component of a novel repair system**	
1203	**Helicase**	**Aful, Mjan, Mthe, Phor, Aper, Aqua, Tmar, Ecoli**	**Helicase, component of a novel repair system**	
2254	**HD-superfamily hydrolase**	**Aful, Mjan, Phor, Aper**	**DNase, component of a novel repair system**	**Operon (partially) conserved in Aper, Mjan, Aful, Phor**
1203	**Superfamily II helicase**	**All archaea, Aqua, Tmar**	**Helicase, component of a novel repair system**	
1857	**Uncharacterized conserved protein**	**All archaea, Aqua**	**Component of a novel repair system**	
1451	**Metal-dependent hydrolase**	**Aful, Mjan, Mthe, Phor, Aqua, Drad, Cace, Ecoli, Nmen, Hpyl, Cjej, Uure**	**Component of a type I restriction-modification system, possibly a metal-dependent protease involved in antiphage response**	**HsdR in Hpyl, Aful, Mthe, Mjan**
0610	**Helicase related to HsdR subunits of Type I restriction-modification systems**	**Aful, Mjan, Mthe, Ecoli, Hinf, Nmen, Hpyl, Cjej, Uure**	**Helicase, component of a novel restriction-modification system**	**HsdS in Aful, Cjej, Mthe**
1743	**Uncharacterized conserved protein**	**Aful, Phor, Aper, Tmar**	**Adenine-N6-specific DNA methylase, component of a novel restriction-modification system**	**Operon (partially) conserved in Tmar, Aful, Phor**
1483	**Uncharacterized conserved protein**	**Aful, Phor, Aper, Tmar**	**AAA+ superfamily ATPase, component of a novel restriction-modification system**	
0553	**SWI2/SNF2 family helicase**		**Helicase/endonuclease, component of a novel restriction-modification system**	
1993	Uncharacterized conserved protein	Mjan, Phor, Aqua, Tmar, Mtub	Function in chromosome condensation	CrcB (membrane protein involved in chromosome condensation) in Mjan, Phor, Aqua, Tmar, Mtub
1909	Uncharacterized conserved protein	All archaea	Transcription factor	DNA-directed RNA polymerase subunits E' and E'' in Aper, Aful,

Table 3. (*Continued*)

				Mjan
0779	Uncharacterized conserved protein	All bacteria except Uure, Mgen, Mpne, Ctra, Cpne	Function in transcription termination	NusA in Bsub, Cace, Ecoli, Hinf, Rpro, Cjej, Drad, Aqua, Tmar
0037	PP-loop superfamily ATPase	All bacteria	PP-loop ATPase involved in cell cycle control/chaperone activity	FtsH in Bs, Tm, Rp, Ct, Cp
1939	Uncharacterized conserved protein	Tm, Bs, Ca, Ssp	Function in translation	CysS, yacO (rRNA methylase) in Bsub, Tmar, Synecho,
1399	Metal-binding, possibly nucleic acid-binding protein	Aquae, Tmar, Drad, Cace, Synecho, Ecoli, Bsub, Mtub, Hinf, Nmen	Function in translation, possibly specialized ribosomal protein	RpmF in Ecoli, Hinf, Nmen, Cace
1837	RNA-binding protein (contains KH-domain)	All bacteria except for Ecoli, Hinf, Nmen, Rpro, Uure, Mgen, Mpne,	RNA-binding protein involved in translation, possibly specialized ribosomal protein	Ffh, rpsP in Bsub, Cace, Cjej, Synecho, Drad, Aqua, Tmar
1385	Uncharacterized conserved protein	All bacteria except Mgen, Mpne, Ctra, Cpne	Function in translation	PrmA (methylase for ribosomal protein L11) in Bsub, Cace, Drad
0618	DHH-family hydrolase	All archaea and bacteria except Mthe, Ecoli, Hinf, Hpyl, Cjej, Ctra, Cpne, Rpro	Function in translation, possibly exopolyphosphatase	InfB, rbfA in Mtub, Cace (long transcription-translation operon)
2890 (new COG)	SAM-dependent methyltransferase	All archaea and bacteria except Aper	rRNA or tRNA methylase	PrfA in Bsub, Mtub, Ecoli, Hinf, Tpal, Mgen, Mpne, Uure, Ctra, Cpne; Fusion with a translation factor (SUA5) in Mgen, Mpne
2888 (new COG)	Uncharacterized proteins	All archaea	RNA-binding protein, function in translation	Translation factor EF1β in Aful, Mjan, Mthe, Aper
2260	RNA-binding protein	All archaea	Function in translation, possibly ribosome-associated protein	Translation factor EIF2α in Aful, Mjan, Mthe, Phor, Aper
2118	Uncharacterized conserved protein	All archaea	Function in translation, possibly ribosome-associated protein	Belongs to a ribosomal operon in Aful, Mjan, Mthe Aper
1534	RNA-binding protein	All archaea, Ecoli, Hinf, Nmen, Bsub, Cace	Function in translation, possibly ribosome-associated protein	Belongs to a ribosomal operon in Aful, Mthe, Aper
2106	Uncharacterized conserved protein	Aful, Mthe, Phor, Aper	Function in translation, possibly ribosome-associated protein	Belongs to ribosomal operon in Aper, Aful, Mthe
2117	PP-loop superfamily ATPase	Mjan, Mthe, Phor	Subunit of tRNA(5-methylaminomethyl-2-thiouridylate) methyltransferase, contains the PP-loop ATPase domain; distantly related to bacterial tRNA(5-methylaminomethyl-2-thiouridylate) methyltransferase subunits	Belongs to ribosomal operon in Mjan, Mthe
1161	GTPase	Mjan, Phor, Aper, Tmar, Bsub, Cace, Synecho, Uure, Mgen, Mpne, Bbur, Nmen	GTPase, specialized translation factor	RplS, rnh in Bsub, Cace RplS, trmD in Mgen, Mpne, Uure
0536	GTPase	All bacteria	GTPase, translation factor	RpmA, rplU in Bsub, Cace, Mtu, Ecoli
0012	GTPase	Ubiquitous	GTPase, translation factor	Pth in Ecoli, Hinf, Nmen, Rpro P14 subunit of Rnase P in Aper L15E in *Sulfolobus solfataricus*
0486	**GTPase**	**All bacteria except Mtub**	**GTPase involved in translation and cell division**	**RpmH, rnpA (RNase P), yidC (preprotein translocase) in Ecoli, Hinf, Bsub, Hpyl, Cjej; GidA, gidB in Bsub, Cace, Bbur**
1847	**RNA-binding protein**	**Tmar, Drad, Bsub, Cace, Synecho, Mtub, Bbur, Tpal**	**RNA-binding protein involved in translation and cell division; possibly a specialized ribosomal protein**	**RnpA, rpmH, thdF (GTPase, COG0486) in Bsub, Cace, Bbur, Tpal YidC in Tmar, Cace, Mtub GidA, gidB in Bsub, Cace**
0759	**Uncharacterized conserved protein**	**ll bacteria except Uure, Mgen, Mpne**	**Function in translation/cell division**	**YidC, jag (COG8147), rpmH, rnpA in Tmar, Drad, Hinf, Nmen, Hpyl, Synecho**
1718	**Ser/Thr protein kinase**	**All archaea, Drad**	**Ser/Thr protein kinase involved in translation regulation**	**Gene doublet conserved in all archaea; IF-1 in Mjan, Phor, Mthe**
1094	**RNA-binding protein**	**All archaea**	**RNA-binding potein involved in translation regulation**	
1491	**RNA-binding protein**	**All archaea**	**Function in translation, possibly ribosome-associated protein**	**Operon with ksgA (dimethlyadenosine transferase) (partially) conserved in all archaea**
1460	**Uncharacterized conserved protein**	**All archaea**	**Function in translation**	
0802	**ATPase or kinase**	**All bacteria except Uure, Mgen, Mpne**	**ATPase or kinase, component of a translation-associated chaperone complex**	**Partially conserved operon in Bsub, Cace, Mtub, Bbur, Tpal, Ecoli, Hinf, Nmen GroES, GroEL in Bsub, Mtub, Cace RpsU in Ecoli, Hinf, Xfas**

(Continues on following page)

Table 3. (*Continued*)

1214	**Inactive homolog of metalloproteases**	**All bacteria**	**Inactive homolog of metalloproteases (Aravind and Koonin 1999), molecular chaperone, component of a translation-associated chaperone complex**	
0456	**Acetyltransferase**	**All archaea and bacteria except Mgen, Mpne, Hpyl, Bbur, Tpal, Cpne, Rpro**	**Acetyltransferase, component of a component of a translation-associated chaperone complex; probably involved in modification of translation machinery components (*E. coli* rimI)**	
0533	**Metalloprotease with possible chaperone activity**	**Ubiquitous**	**Metalloprotease (Aravind and Koonin 1999), molecular chaperone with a central role in translation/replication/ transcription, component of a translation-associated chaperone**	
1236, 2123, 0689, 1096 1097, 0638, 1603, 325, 1369, 1500, 2136, 2042, 1412, 2892 (new COG)		**Typically, all archaea**	**Components of the predicted Archaeal exosome-proteasome system for RNA and protein degradation; (Koonin et al. 2001)**	**3-4 operons partially conserved in all archaea**
0557	Exoribonuclease (vacB)	All bacteria except Mtub, Bbur, Rpro	Exoribonuclease involved in tmRNA degradation?	SmpB (tmRNA-binding protein) in Bsub, Cace, Uure
2739	Uncharacterized conserved protein	Bsub, Cace, Uure, Mpne	Regulator of transcription of the gene for signal recognition particle GTPase (ffh)	Ffh in Bsub, Cace, Uure, Mpne
1585	Uncharacterized membrane protein	All archaea, Aqua, Tmar, Drad, Bsub, Cace, Synecho, Ecoli, Mtub	Membrane protease subunit	Membrane protease subunit (stomatin homolog) in Mtub, Tmar, Mjan, Mthe; Fusion with a ClpP-like protease in Bsub
2001	Uncharacterized conserved protein	Drad, Cace, Bsub, Ecoli, Mtub, Hinf, Nmen, Uure, Mgen, Mpne, Tpal, Rpro	Function in cell division /envelope biogenesis	Envelope biogenesis/cell division operon in Bsub, Cace, Drad, Mtu, Hinf, Nmen
2743	Uncharacterized conserved protein	Bsub, Cace, Mtub	Function in cell division/envelope biogenesis	Envelope biogenesis/cell division operon in Bsub, Cace, Mtu
2744	Uncharacterized conserved protein	Bsub, Cace, Mtub	Function in cell division/envelope biogenesis	Envelope biogenesis/cell division operon in Bsub, Cace, Mtu
0455	MinD-family ATPase	All bacteria and archaea except Aper, Hinf, Uure, Mgen, Mpne, Ctra, Cpne, Rpro	MinD-family ATPase involved in flagellar function/biogenesis	Flagellar operon in Bsub, Cace, Hpyl. Cjej, Aqua, Tmar, Bbur, Tpal
1828	Uncharacterized conserved protein	Af, Mj, Mth, Pa, Aa, Tm, Dr, Ssp, Bs, Cj	Function in purine biosynthesis	PurC, purL, purQ in Bs, Cj, Dr, Tm
0742	N6-adenine-specific methylase	All bacteria except Aa, Mg, Mp, Uu	A methyltransferase involved in Coenzyme A biosynthesis	CoaD (phosphopantetheine adenylyltransferase) in Bsub, Cace, Mtub, Drad
1713	HD superfamily hydrolase	Tmar, Drad, Cace, Synecho, Bsub, Uure, Mgen, Mpne, Tpal	Hydrolase involved in NAD metabolism	NadD in Bsub, Cace, Tpal; Fusion with NadD in Uure, Mgen, Mpne
0799	Uncharacterized conserved protein	All bacteria except Uure, Mgen, Mpne	Enzyme involved in NAD metabolism	NadD, COG1713 in Bsub, Cace, Drad, Tpal
0596[c]	Hydrolases or acyltransferases of the a/b hydrolase superfamily	Bsub, Ecoli, Vcho, Hinf	Hydrolase or acyltransferase involved in menaquinone biosynthesis	MenBDE (menoquinone operon) in Bsub, Ecoli, Vcho, Hinf
1540	Uncharacterized conserved protein	Phor, Drad, Ecoli, Vcho, Hinf, Nmen, Bsub, Cjej	Enzyme of nitrogen/urea metabolism	Allophanate hydrolase subunits in Phor, Vcho, Nmen, Cjej, Drad
1545	Uncharacterized conserved protein	All archaea, Mtub	Function in fatty acid metabolism	Two genes for acetyl-CoA acetyltransferases in Mtub, Aper, Aful; one of the paralogs in Aful fused to 3-oxoacyl-[acyl-carrier-protein] synthase
1327	ATP- and DNA-binding protein	Tmar, Drad, Ecoli, Hinf, Nmen, Bsub, Cace, Synecho, Mtub, Ctra, Cpne	Transcriptional regulator of riboflavine operon	RibD, ribH in Ecoli, Vcho, Hinf, Nmen
0834	Periplasmic amino-acid-binding	All bacteria except Aqua, Uure,	Periplasmic glutamate-binding protein	Glutamate transport operon in Ecoli,

Table 3. (*Continued*)

	protein	Mgen, Mpne, Bbur; Aful, Aper		Hinf
1699	Uncharacterized conserved protein	Tmar, Bsub, Cace, Hpyl, Cjej, Bbur, Tpal	Flagellar subunit	Flagellar operon in Bsub, Cace, Bbur, Tpal
2127	Uncharacterized conserved protein	Drad, Bsub, Cace, Synecho, Ecoli, Nmen, Hpyl, Cjej, Mtub,	Function in protein degradation, chaperone activity?	ClpA in Ecoli, Vcho, Nmen, Xfas, Hpyl, Cjej, Cacc
1805	Uncharacterized membrane protein	Tmar, Synecho, Ecoli, Hinf, Nmen, Ctra, Cpne	Membrane subunit of Na+-translocating NADH-quinone reductase	Na+-translocating NADH-quinone reductase operon in Ecoli, Vcho, Hinf, Tmar
1339	Uncharacterized conserved protein	All archaea	DNA-binding proteins, transcriptional regulator of riboflavin/FAD biosynthesis	3,4-dihydorxy-2-butanone 4-phosphate synthase in Aful, Mjan, Mthe
2061	ACT-domain-containing protein	Aful, Mjan, Mthe	Regulatory subunit of homoserine dehydrogenase	Homoserine dehydrogenase in Aful, Mjan, Mthe
2425	Ucharacterized conserved protein	Mjan, Aper, Ecoli	Mg-binding subunit of Mg-chelatase	ATPase subunit of Mg-chelatase in Mjan, Aper, Ecoli
9714	MoxR-like ATPase	Mjan, Aper, Ecoli	ATPase subunit of Mg-chelatase	Predicted Mg-binding subunit (von Willebrand A domain) of Mg-chelatase in Mjan, Aper, Ecoli
2064	Uncharacterized membrane protein	All archaea	Function in flagellar biosynthesis	ATPase involved in flagellar biosynthesis in all archaea
2822 (new COG)	Uncharacterized proteins	Ecoli, Nmen, Bsub	iron-binding periplasmic lipoprotein	High affinity iron permease, COG2837 in Ecoli, Nmen, Bsub
2837 (new COG)	Uncharacterized proteins	Drad, Ecoli, Nmen, Bsub	iron-dependent peroxidase	High affinity iron permease, COG2822 in Ecoli, Nmen, Bsub

[a]The shaded blocks separated by blank rows indicate predicted novel operons.
[b]From the COG collection as of September 1, 2000.
[c]Some of the predictions involved additional, detailed sequence analysis.
[d]The adjacent genes known to encode proteins with the respective function are indicated.
[e]This prediction only applies to a subset of proteins in this large COG.

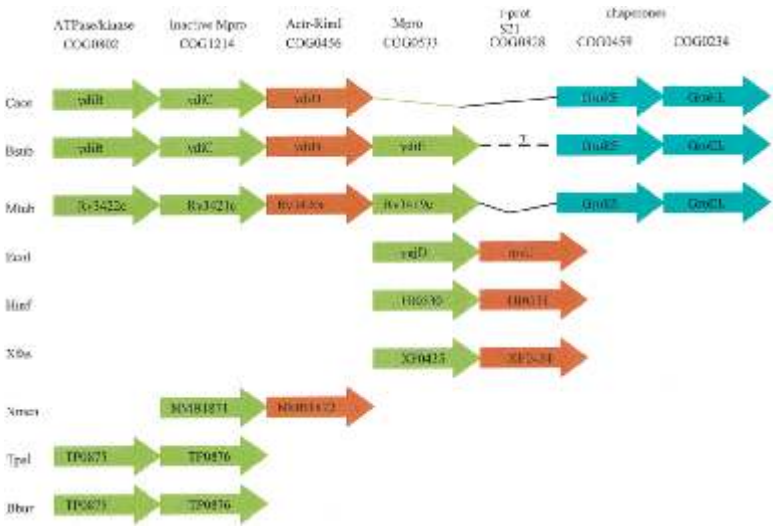


Figure 8 Organization of genes for subunits of the predicted novel, translation-associated chaperone complex in bacterial chromosomes. Genes are shown as arrows (not drawn to scale). Green arrows show genes coding for subunits of the postulated chaperone complex; orange arrows show genes for translation-associated proteins; blue arrows show genes encoding molecular chaperones. The corresponding systematic gene names are indicated. For *Clostridium*, the genes are provisionally named after their *Bacillus* orthologs. *B. subtilis* has a 7-gene insert indicated by a broken line. The (predicted) functions of the gene products and the corresponding COG numbers are indicated above the aligned genes. Abbreviations: Mpro, metalloprotease; Actr, acetyltransferase; r-prot, ribosomal protein.

the latter finding and its multiple implications are presented in detail in Koonin et al. 2001). Some of the functional predictions that result from the analysis of archaeal operons have the extra benefit of the extrapolation to eukaryotic orthologs of archaeal proteins, for which the data on gene order cannot be used directly; the predicted exosome-proteasome operons are the strongest example (Koonin et al. 2001).

As an example of a gene-order-based functional prediction with multiple implications, we discuss here in some detail the predicted, novel translation-linked chaperone complex. Several overlapping configurations of the predicted operon were detected in multiple bacterial genomes (Fig. 8). The core of the operon includes two genes with predicted chaperone functions, namely the metalloprotease with the HSP70 fold (COG0533) and its diverged paralog in which the catalytic site is disrupted (COG1214). It appears likely that both of these proteins, particularly the inactivated version, function as molecular chaperones (Aravind and Koonin 1999). The adjacency of the genes encoding the GroEL-GroES chaperonin–co-chaperonin pair in *Clostridium*, *Mycobacterium*, and

(with an insert) *Bacillus* (Fig. 8) is an additional indication of a chaperone function for the proposed complex. A connection with translation is suggested by the presence of a gene that encodes the RimI acetyltransferase involved in acetylation of ribosomal proteins (Yoshikawa et al. 1987; COG 0456) in *Clostridium*, *Bacillus*, *Mycobacterium*, and *Neisseria* and the gene for the ribosomal protein S21 in Gamma-Proteobacteria (Fig. 8). This case illustrates the complementary contributions of different genomes to functional prediction based on gene order conservation. No single genome contains all the relevant genes within a single string (predicted operon), but the presence of different functionally suggestive gene combinations in different genomes seems to make a strong case for the existence of a previously undetected protein complex with a chaperone-like activity. In *Clostridium*, *Bacillus*, and *Mycobacterium*, the subunits of this predicted complex appear to be encoded by a single operon that seems to have undergone various disruptions and rearrangements in other bacteria.

An additional application of the gene order conservation analysis is the ability to clarify orthologous relationships between genes on the basis of their genomic context. An example of the delineation of a previously unrecognized set of orthologs includes the new COG2890, which consists of predicted rRNA or tRNA methylases that have been previously lumped in a single COG with a variety of other methyltransferases. Furthermore, several new COGs were identified when previously untranslated genes orthologous to genes found in a particular, conserved genomic context were detected, e.g., the new COGs 2888 and 2890.

Conclusions

The systematic comparison of gene order in bacterial and archaeal genomes confirms the notion that there is very little, if any, conservation above the operon level between phylogenetically distant genomes. A corollary is that whenever statistically significant conservation of gene order is observed, it should be considered an indication of operon organization of the respective genes and a legitimate basis for the prediction of functional and potentially physical interactions between genes, which, through 'guilt by association,' helps in predicting functions of uncharacterized genes. However, the same evolutionary force that makes conserved gene strings functionally relevant (namely, intensive intragenomic recombination) limits the utility of the gene order data for functional prediction, because only a minority of the potential operons in any given genome are covered by alignments with sufficiently distant genomes from the current collection of complete genomes. It appears that many more genomes need to be sequenced to significantly increase this coverage. Multiple genomes separated by intermediate evolutionary distances such as representatives of different genera within the same bacterial family could be particularly helpful for making the best use of gene order conservation.

A detailed examination of template-anchored multiple genome alignments for all completely sequenced archaeal and bacterial genomes resulted in new functional predictions that have not been attained previously despite detailed sequence analysis, for ~4% of the ancient conserved protein families represented in the COG collection. Thus, gene order analysis provides for a significant incremental increase in the functional prediction rate for complete prokaryotic genomes, although the contribution of gene order analysis is not comparable to that from direct sequence comparison. These limitations notwithstanding, many potentially important predictions were made, particularly for archaeal genomes. Whereas the 'genomescape' of bacterial genomes appears largely familiar and is dominated by well-characterized operons that encode ribosomal proteins, ABC-type transport cassettes, and enzymes of known metabolic pathways, there are many uncharacterized predicted operons that are conserved in archaeal genomes. Examination of these using a combination of gene order comparison and sequence analysis with sensitive methods suggests previously unsuspected aspects of archaeal biology, which may have implications for the functions of eukaryotic homologs of the respective genes.

A rough correlation exists between the conservation of gene order (number of conserved gene strings) and genome-wide measures of evolutionary distance between genomes, such as the conservation of the gene repertoire (defined as the co-occurrence in families of orthologous proteins) and the median level of similarity between orthologs. However, to a much greater extent than these measures, the fraction of the prokaryotic genomes that belongs to conserved operons varies within a wide range without an obvious pattern of correlation with the phylogenetic relationships or the lifestyles of the respective species. The identification of the functional cognates of these major differences in the level of operon conservation could potentially reveal new aspects of bacterial and archaeal biology.

METHODS

Genome Sequences, Databases, and Sequence Analysis

The annotated genome sequences with the accompanying information on the positions and transcription directions of all protein-coding genes were retrieved from the Genomes division of the Entrez system. The following genomes were analyzed: bacteria — *Aquifex aeolicus* (Aquae), *Bacillus subtilis* (Bsub), *Borrelia burgdorferi* (Bbur), *Campylobacter jejunii* (Cjej), *Chlamydia trachomatis* (Ctra), *Chlamydia pneumoniae* (Cpne),

Clostridium acetobutilicum (Cace), *Deinococcus radiodurans* (Drad), *Escherichia coli* (Ecoli), *Haemophilus influenzae* (Hinf), *Helicobacter pylori* (Hpyl), *Mycoplasma genitalium* (Mgen), *Mycoplasma pneumoniae* (Mpne), *Mycobacterium tuberculosis* (Mtub), *Neisseria meningitidis* (Nmen), *Synechocystis* PCC6803 (Synecho), *Thermotoga maritima* (Tmar), *Treponema pallidum* (Tpal), and *Ureaplasma urealyticum* (Uure), and archaea — *Archaeoglobus fulgidus* (Aful), *Methanobacterium thermoautotrophicum* (Mthe), *Methanococcus jannaschii* (Mjan), *Pyrococcus horikoshii* (Pyro) (*Euryarchaeota*), and *Aeropyrum pernix* (Aero) (*Crenarchaeota*). In addition, the genomes of *Xylella fastidiosa* (Xfas) and *Vibrio cholerae* (Vcho) that have become available during this work were used for the analysis of some of the conserved gene strings.

For the construction of genome alignments, an all-against-all comparison of the protein sequences encoded in the complete prokaryotic genomes was performed using the BLASTP program (Altschul et al. 1997). The alignment scores were calculated as the information density, i.e., the BLAST score expressed in bits divided by the length of the shorter sequence in the alignment. Additional, iterative database searches for detailed analysis of protein sequences were run against the nonredundant database of protein sequences (National Center for Biotechnology Information, NIH, Bethesda, MD) using the PSI-BLAST program (Altschul et al. 1997; Altschul and Koonin 1998). Protein sequences were also compared to the database of Clusters of Orthologous Groups of proteins (COGs; http://www.ncbi.nlm.nih.gov/COG/) using the COGNITOR program (Tatusov et al. 1997, 2000), to the database of domain-specific Hidden Markov Models using the SMART program (Schultz et al. 2000), and to the NCBI's CD (Conserved Domains) collection of position-specific scoring matrices using the reversed PSI-BLAST program (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi).

Local Gene-By-Gene Genome Alignments

For a pair of genomes, the results of all-against-all protein comparisons were postprocessed in two ways, depending on what scoring scheme was used. For the information-density-based scheme, all scores were normalized in such a way that the average score of a unidirectional best hit equaled *1*. For the ortholog-based scheme, the gene pairs that formed a bidirectional best hit were assigned a score of *1*, and all other pairs were assigned a score of *0*. In both cases, the pairs of genes whose protein products did not yield statistically significant local alignments (*E*-value <0.01 with the search space adjusted to the size of the nonredundant database) were assigned a *0* score. The program Lamarck was written for constructing local genome alignments. Lamarck first exhaustively searches for all ungapped local alignments that have at $\geqslant$2 positively scoring matches within a window of length 4 and then heuristically attempts to link the ungapped alignments into longer chains. Scores of these chains were computed using empirically chosen gap opening, gap extension, and spacer penalties of 0.0, 0.3, and 0.3, respectively; a linked chain is accepted if its score is greater than the sum of the scores of the original alignments. Statistical significance of the local alignments was estimated using Monte Carlo simulations. For each pair of genomes, 100 pairs of shuffled gene sequences with the same distribution of BLAST scores as the real pair were aligned, and a score distribution of gene strings was produced, which allows the estimation of the random expectation (*E*) for each score value.

Other Methods of Genome Comparison

The table of co-occurrence of genomes in COGs is available on the COG Web site (http://www.ncbi.nlm.nih.gov/COG/l). The co-occurrence Jackard coefficients were calculated as $Q_{ij} = C_{ij}/(N_i + N_j - C_{ij})$ where C_{ij} is the number of COGs in which genomes *i* and *j* cooccur, and N_i and N_j are the numbers of COGs that include the genomes *i* and *j*, respectively (Sneath and Sokal 1973). Similarly, co-occurrence coefficients for conserved gene strings were calculated from the numbers of genes with positive scores in the pairwise alignment of the genomes *i* and *j* and from the total number of genes from the respective genomes that participate in at least one pairwise alignment with any of the other genomes. The distributions of percent identity between probable orthologs for each pair of genomes were calculated from the results of all-against-all protein comparisons as described previously (Grishin et al. 2000).

Availability of the Complete Results

Template-anchored and pairwise local genome alignments for all completely sequenced bacterial and archaeal genomes are available at ftp://ncbi.nlm.nih.gov/pub/koonin/genome_align. The program Lamarck for local genome alignment is available from the authors upon request.

ACKNOWLEDGMENTS

We thank L. Aravind and Kira Makarova for helpful discussions. The genome sequence of *acetobutilicum* was released by Genome Therapeutic Corporation prior to publication through the NCBI's 'BLAST with Microbial Genomes' Web page (http://www.ncbi.nlm.nih.gov/Microb_blast/unfinishedgenome.html). This prepublication release is gratefully acknowledged.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

NOTE ADDED IN PROOF

After the manuscript of this article was submitted for publication, several new developments in comparative analysis of prokaryotic gene order have been published. A new method for constructing multiple local genome alignments and its application for comparative analysis of 17 complete archaeal and bacterial genomes has been reported (Fujibuchi et al. 2000). Some of the conclusions reached by Fujibuchi and coworkers overlap with the results presented in the present article, particularly with respect to the coverage of different genomes with conserved gene clusters. The STRING web server for detection of conserved gene clusters in multiple genomes has been described (Snel et al. 2000). In addition, an analysis of the genome of *M. genitalium* benchmarking the context approach to gene function prediction has been published (Huynen et al. 2000b). New functional predictions were reported for ~10% of *M. genitalium* genes which is a greater rate than we report in the present paper. This is probably due to a somewhat less restrictive approach to the 'novelty' of predictions employed by these workers and also to the fact that we limited our analysis to evolutionarily conserved proteins included in the COGs. Also after this article was submitted for publication, the COG database was significantly updated (Tatusov et al. 2001). However, the COG numbers remain stable,

so they still can be used to access any of the COGs mentioned in this article.

REFERENCES

Altschul, S.F. 1998. Generalized affine gap costs for protein sequence alignment. *Proteins* **32:** 88–96.

Altschul, S.F. and Gish, W. 1996. Local alignment statistics. *Methods Enzymol.* **266:** 460–480.

Altschul, S.F. and Koonin, E.V. 1998. PSI-BLAST — a tool for making discoveries in sequence databases. *Trends Biochem. Sci.* **23:** 444–447.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Aravind, L. and Koonin, E.V. 1999. Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches. *J. Mol. Biol.* **287:** 1023–1040.

Dandekar, T., Snel, B., Huynen, M., and Bork, P. 1998. Conservation of gene order: A fingerprint of proteins that physically interact. *Trends Biochem. Sci.* **23:** 324–328.

Fitch, W.M. 1970. Distinguishing homologous from analogous proteins. *Syst. Zool.* **19:** 99–106.

Fujibuchi, W., Ogata, H., Matsuda, H., and Kanehisa, M. 2000. Automatic detection of conserved gene clusters in multiple genomes by graph comparison and P-quasi grouping. *Nucleic Acids Res.* **28:** 4029–4036.

Galperin, M.Y. and Koonin, E.V. 2000. Who's your neighbor? New computational approaches for functional genomics. *Nat. Biotechnol.* **18:** 609–613.

Glansdorff, N. 1999. On the origin of operons and their possible role in evolution toward thermophily. *J. Mol. Evol.* **49:** 432–438.

Grishin, N.V., Wolf, Y I., and Koonin, E.V. 2000. From complete genomes to measures of substitution rate variability within and between proteins. *Genome Res.* **10:** 991–1000.

Henikoff, S. and Henikoff, J.G. 2000. Amino acid substitution matrices. *Adv. Protein. Chem.* **54:** 73–97.

Himmelreich, R., Plagens, H., Hilbert, H., Reiner, B., and Herrmann, R. 1997. Comparative analysis of the genomes of the bacteria *Mycoplasma pneumoniae* and *Mycoplasma genitalium*. *Nucleic Acids Res.* **25:** 701–712.

Huynen, M.J. and Snel, B. 2000. Gene and context: Integrative approaches to genome analysis. *Adv. Prot. Chem.* **54:** 345–379.

Huynen, M., Snel, B., Lathe, W., and Bork, P. 2000a. Exploitation of gene context. *Curr. Opin. Struct. Biol.* **10:** 366–370.

———. 2000b. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res.* **10:** 1204–1210.

Itoh, T., Takemoto, K., Mori, H., and Gojobori, T. 1999. Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. *Mol. Biol. Evol.* **16:** 332–346.

Jacob, F., Perrin, D., Sanchez, C., and Monod, J. 1960. L'Operon: Groupe de genes a expression coordonee par un operateur. *C.R. Seance Acad. Sci.* **250:** 1727–1729.

Kanehisa, M. and Goto, S. 2000. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28:** 27–30.

Koonin, E.V., Wolf, Y.I., and Aravind, L. 2001. Prediction of the archaeal exosome and its connecitons with the proteasome and the translation and transcription machineries by a comparative-genomic approach. *Genome Res.* **11:** 240–252.

Koonin, E.V., Mushegian, A.R., and Rudd, K.E. 1996. Sequencing and analysis of bacterial genomes. *Curr. Biol.* **6:** 404–416.

Lawrence, J. 1999. Selfish operons: The evolutionary impact of gene clustering in prokaryotes and eukaryotes. *Curr. Opin. Genet. Dev.* **9:** 642–648.

———. 1997. Selfish operons and speciation by gene transfer. *Trends Microbiol.* **5:** 355–359.

Martin, B., Garcia, P., Castanie, M.P., Glise, B., and Claverys, J.P. 1995. The recA gene of *Streptococcus pneumoniae* is part of a competence-induced operon and controls an SOS regulon. *Dev. Biol. Stand.* **85:** 293–300.

Mushegian, A.R. and Koonin, E.V. 1996. Gene order is not conserved in bacterial evolution. *Trends Genet.* **12:** 289–290.

Narumi, I., Satoh, K., Kikuchi, M., Funayama, T., Kitayama, S., Yanagisawa, T., Watanabe, H., and Yamamoto, K. 1999. Molecular analysis of the Deinococcus radiodurans recA locus and identification of a mutation site in a DNA repair-deficient mutant, rec30. *Mutat. Res.* **435:** 233–243.

Needleman, S.B. and Wunsch, C.D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48:** 443–453.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D., and Maltsev, N. 1998. Use of contiguity on the chromosome to predict functional coupling. *In Silico Biol.* 1: 93–108.

———. 1999. The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci.* **96:** 2896–2901.

Salgado, H., Moreno-Hagelsieb, G., Smith, T.F., and Collado-Vides, J. 2000. Operons in *Escherichia coli*: Genomic analyses and predictions. *Proc. Natl. Acad. Sci.* **97:** 6652–6657.

Schultz, J., Copley, R.R., Doerks, T., Ponting, C.P., and Bork, P. 2000. SMART: A web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.* **28:** 231–234.

Siefert, J.L., Martin, K.A., Abdi, F., Widger, W.R., and Fox, G.E. 1997. Conserved gene clusters in bacterial genomes provide further support for the primacy of RNA. *J. Mol. Evol.* **45:** 467–472.

Smith, T.F. and Waterman, M.S. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* **147:** 195–197.

Sneath, P.H.A. and Sokal, R.R. 1973. *Numerical taxonomy*. W.H. Freeman, San Francisco, CA.

Snel, B., Lehmann, G., Bork, P., and Huynen, M.A. 2000. STRING: A web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res.* **28:** 3442–3444.

Spitzer, E.D. and Weiss, B. 1985. dfp gene of *Escherichia coli* K-12, a locus affecting DNA synthesis, codes for a flavoprotein. *J. Bacteriol.* **164:** 994–1003.

Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D., and Koonin, E.V. 2001. The COG database: New developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* **29:** 22–28.

Tatusov, R.L., Mushegian, A.R., Bork, P., Brown, N.P., Hayes, W.S., Borodovsky, M., Rudd, K.E., and Koonin, E.V. 1996. Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli*. *Curr. Biol.* **6:** 279–291.

Tatusov, R.L., Koonin, E.V., and Lipman, D.J. 1997. A genomic perspective on protein families. *Science **278:*** **631–637.**

Tatusov, R.L., Galperin, M.Y., Natale, D.A., and Koonin, E.V. 2000. The COG database: A tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28:** 33–36.

Tomii, K. and Kanehisa, M. 1998. A comparative analysis of ABC transporters in complete microbial genomes. *Genome Res.* **8:** 1048–1059.

Vingron, M. and Waterman, M.S. 1994. Sequence alignment and penalty choice. Review of concepts, case studies and implications. *J. Mol. Biol.* **235:** 1–12.

Watanabe, H., Mori, H., Itoh, T., and Gojobori, T. 1997. Genome plasticity as a paradigm of eubacteria evolution. *J. Mol. Evol.* **44:** S57–64.

Wolf, Y.I., Aravind, L., Grishin, N.V., and Koonin, E.V. 1999. Evolution of aminoacyl-tRNA synthetases — analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. *Genome Res.* **9:** 689–710.

Yoshikawa, A., Isono, S., Sheback, A., and Isono, K. 1987. Cloning and nucleotide sequencing of the genes rimI and rimJ which encode enzymes acetylating ribosomal proteins S18 and S5 of *Escherichia coli* K12. *Mol. Gen. Genet.* **209:** 481–488.

Received August 23, 2000; accepted in revised form December 13, 2000.