Published online October 11, 2004

5392–5397 *Nucleic Acids Research, 2004, Vol. 32, No. 18*
doi:10.1093/nar/gkh882

Conservation of adjacency as evidence of paralogous operons

Sarath Chandra Janga and Gabriel Moreno-Hagelsieb*

Program of Computational Genomics, CIFN-UNAM, Apdo Postal 565-A, Cuernavaca, Morelos, 62100 Mexico

Received June 12, 2004; Revised August 26, 2004; Accepted September 21, 2004

ABSTRACT

Most of the analyses on the conservation of gene order are limited to orthologous genes. However, the organization of genes into operons might also result in the conservation of gene order of paralogous genes. Thus, we sought computational evidence that conservation of gene order of paralogous genes represents another level of conservation of genes in operons. We found that pairs of genes within experimentally characterized operons of *Escherichia coli* K12 and *Bacillus subtilis* tend to have more adjacently conserved paralogs than pairs of genes at transcription unit boundaries. The fraction of same strand gene pairs corresponding to conserved paralogs averages 0.07 with a maximum of 0.22 in *Borrelia burgdorferi*. The use of evidence from the conservation of adjacency of paralogous genes can improve the prediction of operons in *E.coli* K12 by ∼0.27 over predictions using conservation of adjacency of orthologous genes alone.

INTRODUCTION

The infrequent conservation of gene order in prokaryotic evolution (1) has lead to the use of conserved gene clusters as an indication of functional relationships (2,3). Analyses on the conservation of gene order and operon organization have shown that most of the conserved gene order in evolutionarily distant genomes is due to operon organization (4,5). So far, most of the studies about conservation of gene order are focused on orthologs, defined as genes that have diverged after speciation events (6,7).

According to the selfish operon theory, genes with related functions organized into operons would be more successful in horizontal gene transfer events because they would more probably be transferred together and confer a complete function to the new host (8). Genes in operons would also have advantages if kept together during recombination events (9). Sometimes, the result of such events would be the appearance of duplicated or paralogous operons. Two specific examples of operon paralogs are a paper on redundant copies of tryptophan-pathway genes probably organized into operons (10), and another on subtilisin operons in treponemes (11). Genomic works include an analysis of lineage-specific gene duplications where the authors mention the presence of a few paralogous operons (12). In another, Babu and Teichmann (13) found duplicated sets of genes in *Escherichia coli* where each set includes a gene coding for a transcription factor and of the genes they regulate, some of them in operons. Finally, a very recent work showed that several stretches of duplicated genes correspond to known operons (14), but does not offer further quantification.

Here, we study the conservation of adjacency of paralogous genes. We propose the terms intraparalogs for paralogous genes within a given genome, and extraparalogs for those occurring in different genomes. For example, the genes coding for CRP and FNR, both in *E.coli* K12, would be intraparalogs, while the gene coding for CRP in *E.coli* K12 and that coding for FNR in *Haemophilus influenzae* would be extraparalogs. Note that extraparalogs do not require both genomes to contain all sets of genes. These two terms should not be mistaken for the terms introduced by Sonnhammer and Koonin (15), who define inparalogs as those that result from a duplication event occurring after speciation (lineage-specific paralogs), and outparalogs as those resulting from a before-speciation duplication event. In both the cases, Sonnhammer and Koonin refer to distinctions among intraparalogs. Our work starts with a comparison of the conservation of adjacency of intraparalogs to genes within experimentally characterized operons of *E.coli* K12 and *Bacillus subtilis* against that of genes at transcription unit boundaries (TUB). We also estimate the proportion of intraparalog conservation of gene order across prokaryotic genomes and offer a method to using the conservation of adjacency of any homologous genes, orthologs and extraparalogs, to predict operons.

DATA PREPARATION

We ran BLASTP program (16) for comparing all the proteins annotated within the current collection of more than 150 prokaryotic genomes at the Entrez Genome Database (17) (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). We used an *E*-value cutoff of 1e−6 with a database size fixed at 5e+8 (−*z* 5e+8), soft filtering of low information content sequences (the −F 'm S' option of the NCBI BLASTPGP program) and the final Smith–Waterman alignment (−s T) (18). We also required coverage of at least 60% of any of the protein sequences in the alignment. Our working definition of orthology consisted of

*To whom correspondence should be addressed at present address: Department of Biology, Wilfrid Laurier University, 75 University Avenue West, Waterloo, ON, Canada N2L 3C5. Tel: +1 519 884 0710 ext. 2364; Fax: +1 519 884 0464; Email: gmoreno@wlu.ca

BLASTP reciprocal best hits as described elsewhere (19), except for the BLASTP options as outlined above.

Intraparalogs are easily found by the identification of homologs, defined as BLASTP hits, inside each genome. Extraparalogs could be any homolog, found in other genomes, left after removing the orthologs. Going back to our example above, we would detect, at the protein level, that the *fnr* and *crp* genes of *H.influenzae* are both homologous to *fnr* of *E.coli*. We would also find that the reciprocal best hit to *fnr* of *E.coli* is *fnr* in *H.influenzae*, thus *crp* of *H.influenzae* would be an extraparalog to *fnr* of *E.coli*. We took care not to count any adjacently conserved pair more than once.

We used the current data set of transcription units of *E.coli* K12 (20) found in RegulonDB (21,22), and the data set of genes in operons of *B.subtilis* (23) from Itoh *et al*. (24) to built data sets of pairs of genes within operons (WO pairs) and at TUB pairs as explained previously (25). The current data sets contain 713 WO pairs and 429 TUB pairs in *E.coli* K12, and 309 WO pairs and 129 TUB pairs in *B.subtilis*. Some analyses were complemented by the use of the SUPERFAMILY domains database (26,27) to search for particular domains in different genes that might be found together (fused) in another gene. This is somewhat related to previous proposals to predict functional interactions by finding gene fusions (28,29), except that such previous work was based on ortholog analyses and BLASTP results rather than on domain analysis.

Data sets of genes in experimentally determined operons contain examples of intraparalog operons

In *E.coli* K12, WO pairs contain 95 intraparalog pairs, accounting for ∼13.32% of all WO pairs. TUB pairs have four intraparalog pairs (0.93%). In *B.subtilis*, we detected 37 intraparalog WO pairs (11.97%) and no intraparalog TUB pairs. We can leave a single representative pair per conserved family of paralog pairs. This reduces the number of WO pairs to 653 in *E.coli* K12 and 288 in *B.subtilis*. Using these non-redundant sets, we calculated the total possible intraparalog pair conservation (TPIP), consisting of the total number of pairs possible from the total number of paralogs to each gene in all pairs where both genes have paralogs elsewhere. For instance, for an imaginary pair of genes ‘a/b’, gene ‘a’ has a total of five intraparalogs while gene ‘b’ has seven. The maximum possible conserved intraparalog pairs would be five. We did the same count for all pairs and added the values to obtain the TPIP. We then found the adjacently conserved intraparalog pairs (CIPs) and obtained the fraction CIP/TPIP. The fraction of conserved pairs is always higher for WO pairs than for TUB pairs: 172/545 (0.315) and 7/255 (0.027), respectively for *E.coli* K12; 84/248 (0.338) and 2/52 (0.038), respectively for *B.subtilis*. Thus, genes WO have a clear and strong tendency to have adjacently conserved intraparalog genes when compared with genes at TU boundaries.

Most adjacently conserved intraparalogs are organized in operons

From the analyses above, most adjacently conserved intraparalog genes detected should be in operons, at least for *E.coli* K12 and *B.subtilis*. In Figure 1, we show intergenic distance distributions for conserved intraparalog genes as found in different groups of prokaryotes. The distributions confirm the expectation that most conserved intraparalog genes correspond to genes in operons in other organisms.

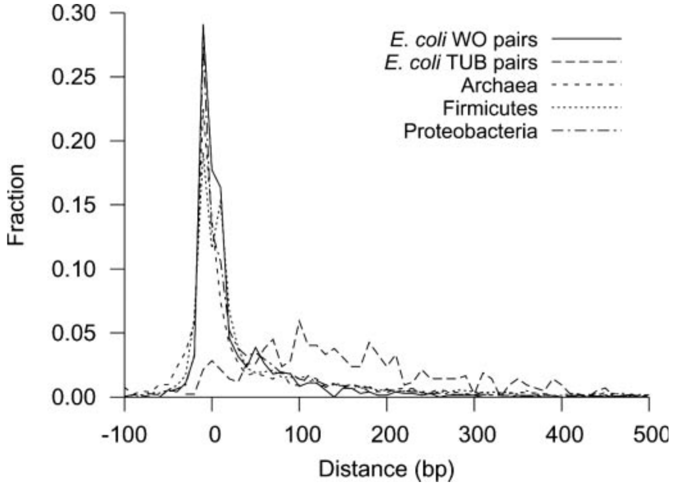

Figure 1. Intergenic distance distribution of adjacently conserved intraparalogs. Note that the distributions resemble that of pairs of genes known to be in operons in *E.coli* K12.

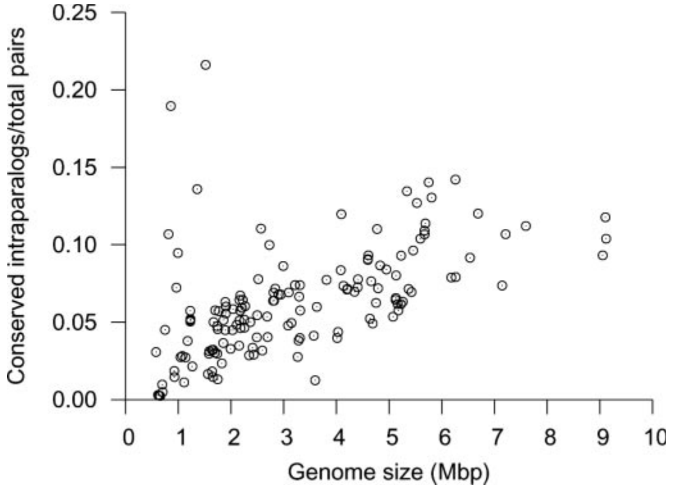

Figure 2. Proportion of adjacently conserved intraparalog pairs of genes. The higher proportions of CIPs of genes per same strand pair of genes occur in two organisms with small genomes: *B.burgdorferi*, with a genome of ∼1.52 Mb (including the size of all its sequenced replicons) and a proportion of nearly 0.22, and *P.asteris* (Onion yellows phytoplasma) with a genome of 0.86 Mb and a proportion of ∼0.19.

The proportion of adjacently conserved intraparalog pairs slightly increases with genome size (Figure 2). However, the higher proportions occur in two organisms with small genomes: (i) *Borrelia burgdorferi* (30), with a genome of ∼1.52 Mb (including the size of all its sequenced replicons) and a proportion of nearly 0.22 intraparalogs per same strand pair of genes; and (ii) *Phytoplasma asteris* (Onion yellows phytoplasma) (31) with 0.86 Mb and a proportion of 0.19 CIPs. Another interesting example is *Mycoplasma penetrans* (32) (1.36 Mb and 0.14 intraparalog conserved pairs). Most other genomes with proportions >0.11 have more than 4.1 Mb.

In *B.burgdorferi*, the most abundant intraparalog conserved pairs are formed with genes annotated as 'conserved hypothetical protein', and none of them has domain assignments in the SUPERFAMILY database. In *Phytoplasma*, the highly conserved pairs are either sigma factors or DNA-binding proteins. In *M.penetrans*, the highly conserved pairs of genes are all P35 lipoprotein homologs, known to increase the antigenic diversity of these organisms (33).

Extraparalog conservation of adjacency enriches operon predictions

A group at The Institute for Genomic Research (TIGR) has developed a method to predict operons on the basis of conservation of adjacency of orthologous genes (4). The authors calculate a confidence value for two adjacent genes in the same strand to be in the same operon based on the comparison of the conservation of adjacent genes in the same strand against the conservation of adjacent genes in opposite strands. The conservation of genes in opposite strands represents random conservation of adjacency. They have currently simplified their formula to:

$$C = 1 - 0.5 \times pf/pr$$

Here, C is the confidence for two genes to be in the same operon. 0.5 is a prior probability for the genes to be in different transcription units. Instead of calculating confidence values for adjacent genes in the same strand, the analysis now expands to the conservation of genes in the same directon, a set of adjacent genes in the same strand with no intervening gene in the opposite strand (25). *Pf* is a 'false probability', meaning the count of pairs of orthologs conserved within 'false directons' (one gene in half one directon, the other gene in half the other directon), and *Pr* is a 'real probability' or the count of orthologs conserved within the same directon in two genomes (D. Maria Ermolaeva, personal communication).

We took an intermediate step between the first TIGR procedure and the current one. Our *Pf* is the count of adjacently conserved pairs of orthologs in opposite strands (convergently and divergently transcribed genes) normalized against the overall proportion of adjacent genes in opposite strands. Our *Pr* is the count of adjacently conserved pairs of orthologs in the same strand normalized against the proportion of adjacent genes in the same strand. Instead of a genome to genome comparison, we performed a one-to-many comparison by grouping other genomes by their similarity scores as calculated from the perspective of the problem genome. Thus, for the problem genome, say *E.coli* K12, we counted the number of pairs of adjacent genes conserved adjacent in any other genome within a given similarity score range, say all the genomes with genomic similarity scores against *E.coli* K12 that range from 0.1 to <0.2. As expected, the confidence in assigning genes to the same operon increases with lower genomic similarity scores (Figure 3).

If we take the results at confidence values >0.95, we predict 516 pairs of genes to be in operons in *E.coli* K12, with 327 of them found among the known WO pairs and 21 that appear within the TUB pairs. These highly conserved TUB pairs have appeared before (4,5). It seems like some of these conflicting pairs are in operons in other genomes (5), or might belong to yet to be discovered complex cases where they can be transcribed as part of the previous transcription unit or start a new one depending on the promoter used (as of this writing, there are 55 pairs of genes of this kind in RegulonDB). Ten of these pairs have both their genes classified in the same Riley functional category (34,35), and four code for proteins with domains fused elsewhere, reinforcing the idea that they might be associated into operons (see Table 1).

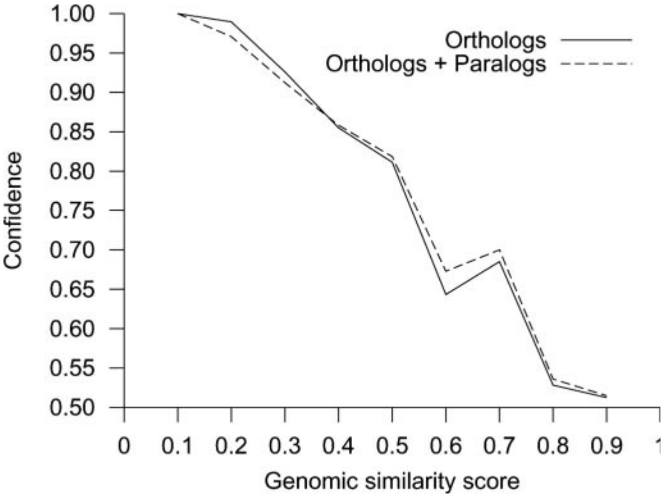

Figure 3. Confidence in operon prediction versus genomic similarity scores. The genomic similarity score is calculated as the sum of all the BLAST bit scores of all putatively ortholog genes between two genomes (comparison score) divided by the sum of the BLAST bit scores of the genes having orthologs in the other genome when compared with themselves (self scores) (19). Note that lower similarity scores, corresponding to evolutionarily more distant genomes, result in a higher confidence that conserved pairs of genes are in the same operon. The confidence values remain about the same when adding extraparalogs to the count of adjacently conserved pairs.

As an assay for the use of paralog conservation of adjacency to predict operons, we used the same method as above, except that we counted all adjacently conserved homologous pairs of genes regardless of them being defined as orthologs or not. By this method, the number of predicted pairs in operons increased to 655 (an increase of ~0.27). Among them, 387 correspond to known WO pairs and 27 to TUB pairs. Most of the false-positive pairs have related functions. Computationally, we find that 12 of these TUB pairs have both their genes classified in the same Riley functional category (34,35), while six code for protein products containing domains that are fused elsewhere (see Table 1).

The functional relationships of 10 of these TUB pairs are evident from their gene names: pairs *lpxD*/*fabZ* and *fabZ*/*lpxA* have *fabZ* in common, making it a cluster of three genes whose products are involved in lipid biosynthesis (36). The protein products of the pair *kdpD*/*kdpC* are involved in K^+ uptake in *E.coli* (37). Pair *pflA*/*pflB* is involved in Energy Metabolisms: Carbon, anaerobic Respiration. Pairs *narK*/*narG* and *narZ*/*narU* are both involved in anaerobic respiration and nitrate/nitrite transport and metabolism (38). Pairs *rplN*/*rpsQ* (in the same operon in *B.subtilis* (24)), *rbsK*/*rbsR* and *rplA*/*rplJ* all code for ribosomal proteins. Finally, genes in pair *hsdM*/*hsdR* code for part of a restriction/modification system (39).

Other gene pairs do not consist of genes that share their nomenclatures yet are functionally related. The genes in pair *tsf*/*pyrH* are part of the common cluster *tsf-pyrH-frr*, where the

Table 1. *E.coli* K12 false positives in operon predictions by homolog (ortholog or paralog) conservation of adjacency

Gene pair	TIGR	Riley	Domain fusions
tsf/pyrH[a]	+	2.3/1.7	—
lpxD/fabZ[a]	+	1.6/1.5	—
fabZ/lpxA[a]	+	1.5/1.6	—
queA/tgt[a]	+	2.2/2.2	—
tig/clpP[a]	–	2.3/1.2; 3.1; 5.5	—
clpX/lon[a]	+	1.2; 2.3/1.2; 3.1	52540, 81296 52540, 57716
kdpD/kdpC[a]	+	2.3; 3.1/4.3	—
yljA(16128849)/*clpA*[a]	+	—	52540, 54736
pflA/pflB[a]	+	1.3; 2.3; 3.1/1.1; 1.3; 1.7	—
plsX/fabH[a]	+	1.6/1.5	—
phoP/purB	–	2.2; 3.1; 3.3/1.5	—
narK/narG[a]	–	1.8; 4.2/1.3; 1.4	—
narZ/narU	–	1.3; 1.4/1.8; 4.2	—
hdhA/malI	–	1.7/1.1; 2.2; 3.1; 3.3	—
nlpC/btuD	–	—	—
tar/chew	–	3.1/3.1	—
rfbB/galF	–	1.5; 1.6; 1.7/1.7	51735, 53448
ackA/pta[a]	+	1.1; 1.3; 1.7/1.1; 1.3; 1.7	52540, 53067
recA/ygaD(16130607)[a]	+	—	—
yhbG(16131091)/*rpoN*[a]	+	4.3/1.8; 2.2; 3.1; 3.3	—
fmt/sun[a]	+	2.2/2.2	—
rplN/rpsQ[a]	+	2.3/2.3	—
gyrB/recF[a]	–	2.1;2.2;3.1/2.1	—
rbsK/rbsR[a]	–	1.1/1.1; 2.2; 3.1; 3.3	47413, 53613
nusG/rplK[a]	+	2.2/2.3	—
rplA/rplJ[a]	–	2.3; 3.1/2.3	—
hsdM/hsdR[a]	–	2.1; 3.1/1.2; 2.1	52540, 53335

[a]Pair also detected by ortholog conservation of adjacency. The names of hypothetical or putative genes are followed by their GI numbers as found in the *E.coli* K12 GenBank file (version: NC_000913.1 GI:16127994). Fifteen of the pairs of genes we predicted were also predicted as genes in operons by Ermolaeva *et al*. (4). Thirteen predictions can be justified as genes that have related or interdependent functions because they either share their Riley classification (34,35), or they contain domains fused elsewhere. Riley functions are as follows: 1.1, Carbon compound utilization; 1.2, Macromolecule degradation; 1.3, Energy metabolism (carbon); 1.4, Energy production/transport; 1.5, Building block biosynthesis; 1.6, Macromolecules (cellular constituent) biosynthesis; 1.7, Central intermediary metabolism; 1.8, Metabolism of other compounds; 2.1, DNA related; 2.2, RNA related; 2.3, Protein related; 3.1, Type of regulation; 3.3, Genetic unit regulated; 4.2, Electrochemical potential driven transporters; 4.3, Primary Active Transporters; 5.5, Adaptation to stress. The domains are (26,27) as follows: 47413, lambda repressor-like DNA-binding domains; 51735, NAD(P)-binding Rossmann-fold domains; 52540, P-loop containing nucleotide triphosphate hydrolases; 53067, Actin-like ATPase domain; 53335, *s*-adenosyl-l-methionine-dependent methyltransferases; 53448, Nucleotide-diphospho-sugar transferases; 53613, Ribokinase-like; 53822, Periplasmic binding protein-like I; 54736, ClpS-like; 57716, Glucocorticoid receptor-like (DNA-binding domain); 81296, E set domains. The ribosomal genes *rplN* and *rpsQ* are known to be in the same operon in *B.subtilis* (24).

product of *pyrH* is involved in nucleotide biosynthesis, while the products of *tsf* and *frr* are involved in translation (40,41). The conservation might be accounted for not from a direct functional relationship, but from the general importance of macromolecular biosynthesis. The pair of genes *queA/tgt* is related to tRNA biosynthesis and modification. In *tig/clpP*, the product of *tig* is a chaperone and that of *clpP* is a heat shock protein (42). In *clpX/lon*, the first gene is related to heat shock, and both genes in the pair are proteases (42). Genes in *plsX/fabH* code both for proteins involved in phospholipid biosynthesis, though not yet annotated as in the same operon in RegulonDB, the genes have been found to be co-transcribed in *E.coli* K12 (43). The genes in *phoP/purB* have also been found to be in the same operon in *E.coli* K12 (44). Genes in pair *tar/cheW* code for proteins involved in chemotaxis (45). In *rfbB/galF*, both genes are involved in the biosynthesis of lipopolysaccharides (46,47). The genes in pair *ackA/pta* code for enzymes in the Embden–Meyerhof–Parnas pathway, phosphotransacetylase (pta) and acetate kinase (ackA) (48), and have been found to be part of the same operon in *E.coli* (49). The genes in *gyrB/recF* are involved in DNA structure and organization. In pair *nusG/rplK*, *nusG* encodes a 181 amino acid long polypeptide and is involved in transcription antitermination (50), and *rplK* codes for a ribosomal protein. The gene *nusG* is in an operon with gene *secE*, coding for a protein export factor, and the *secE-nusG* operon is between the *tufB* and *rplK* genes, both involved in translation. This is similar to the *tsf-pyrH-frr* situation described above, where the flanking genes code for proteins involved in translation, and the genes in the middle have perhaps unrelated, yet important, functions in macromolecular biosynthesis. On the unrelated side, in the pair *nlpC/btuD* the product of the first gene is related to a family of cell wall peptidases (51) and that of the latter to the transport of vitamin B12, the products of these two genes do not seem to be obviously related except by being membrane-bound proteins. The pair *hdhA/malI* does not seem to be justified either, as one is the 7 alpha-hydroxysteroid dehydrogenase (52) and the other is a repressor for the *malX* and *malY* genes. The two genes have no further relationship than an NAD(P)-binding Rossmann-fold.

Except for two (or four if we do not admit the macromolecular biosynthesis examples, *tsf/pyrH* and *nusG/rplK*, as functionally related), all false-positive pairs described above have a functional relationship. Thus, it holds true that most of the conservation of gene order in evolutionarily distant prokaryotes is due to operon organization (4,5). In the few cases where the conservation is not related to operon organization the conserved genes have a functional relationship. The four remaining pairs, *yljA/clpA*, *recA/ygaD*, *yhbG/rpoN* and *fmt/sun*, each contain one hypothetical gene. It will be interesting to explore the functional relationships that these pairs might have.

CONCLUSIONS

The present work shows that conservation of adjacency of homolog genes is related to operon organization, regardless of whether they correspond to orthologs or not. One of the reasons for this conservation might be that the biological re-usability of successful modules occurs at many levels, from protein folding motifs to complete sets of biochemical reactions. Further work will be necessary to test this hypothesis. Our results open up the possibility of complementing current predictions of operons based on conservation of adjacency to the use of any homolog conservation, thus resulting in a higher number of predictions of functional interactions for use in genomic context analyses. This is especially important if a researcher wants to use information from the number of non-closed genomic sequences that will be inundating the databases (53,54), where adjacently conserved genes might not necessarily correspond to orthologs (the true orthologs might not have been sequenced), but would be helpful for operon predictions anyway.

ACKNOWLEDGEMENTS

We thank Warren Lamboy for critical reading of the manuscript. We also acknowledge Victor del Moral, Romualdo Zayas and Cesar Bonavides for computer support. S.C.J. has been supported by grants given to Julio Collado-Vides.

REFERENCES

1. Mushegian,A.R. and Koonin,E.V. (1996) Gene order is not conserved in bacterial evolution. *Trends Genet.*, **12**, 289–290.
2. 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.
3. Overbeek,R., Fonstein,M., D'Souza,M., Pusch,G.D. and Maltsev,N. (1999) The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA*, **96**, 2896–2901.
4. Ermolaeva,M.D., White,O. and Salzberg,S.L. (2001) Prediction of operons in microbial genomes. *Nucleic Acids Res.*, **29**, 1216–1221.
5. Moreno-Hagelsieb,G., Trevino,V., Perez-Rueda,E., Smith,T.F. and Collado-Vides,J. (2001) Transcription unit conservation in the three domains of life: a perspective from *Escherichia coli*. *Trends Genet.*, **17**, 175–177.
6. Fitch,W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.*, **19**, 99–113.
7. Fitch,W.M. (2000) Homology a personal view on some of the problems. *Trends Genet.*, **16**, 227–231.
8. Lawrence,J.G. and Roth,J.R. (1996) Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics*, **143**, 1843–1860.
9. Lawrence,J. (1999) Selfish operons: the evolutionary impact of gene clustering in prokaryotes and eukaryotes. *Curr. Opin. Genet. Dev.*, **9**, 642–648.
10. Xie,G., Bonner,C., Brettin,T., Gottardo,R., Keyhani,N. and Jensen,R. (2003) Lateral gene transfer and ancient paralogy of operons containing redundant copies of tryptophan-pathway genes in Xylella species and in heterocystous cyanobacteria. *Genome Biol.*, **4**, R14.
11. Correia,F.F., Plummer,A.R., Ellen,R.P., Wyss,C., Boches,S.K., Galvin,J.L., Paster,B.J. and Dewhirst,F.E. (2003) Two paralogous families of a two-gene subtilisin operon are widely distributed in oral treponemes. *J. Bacteriol.*, **185**, 6860–6869.
12. Jordan,I.K., Makarova,K.S., Spouge,J.L., Wolf,Y.I. and Koonin,E.V. (2001) Lineage-specific gene expansions in bacterial and archaeal genomes. *Genome Res.*, **11**, 555–565.
13. Madan Babu,M. and Teichmann,S.A. (2003) Evolution of transcription factors and the gene regulatory network in *Escherichia coli*. *Nucleic Acids Res.*, **31**, 1234–1244.
14. Gevers,D., Vandepoele,K., Simillion,C. and Van de Peer,Y. (2004) Gene duplication and biased functional retention of paralogs in bacterial genomes. *Trends Microbiol.*, **12**, 148–154.
15. Sonnhammer,E.L. and Koonin,E.V. (2002) Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet.*, **18**, 619–620.
16. 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.
17. Tatusova,T.A., Karsch-Mizrachi,I. and Ostell,J.A. (1999) Complete genomes in WWW Entrez: data representation and analysis. *Bioinformatics*, **15**, 536–543.
18. Schaffer,A.A., Aravind,L., Madden,T.L., Shavirin,S., Spouge,J.L., Wolf,Y.I., Koonin,E.V. and Altschul,S.F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.*, **29**, 2994–3005.
19. Moreno-Hagelsieb,G. and Collado-Vides,J. (2002) A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics*, **18** (Suppl. 1), S329–336.
20. Blattner,F.R., Plunkett,G.,III, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F. *et al.* (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1474.
21. Huerta,A.M., Salgado,H., Thieffry,D. and Collado-Vides,J. (1998) RegulonDB: a database on transcriptional regulation in *Escherichia coli*. *Nucleic Acids Res.*, **26**, 55–59.
22. Salgado,H., Gama-Castro,S., Martinez-Antonio,A., Diaz-Peredo,E., Sanchez-Solano,F., Peralta-Gil,M., Garcia-Alonso,D., Jimenez-Jacinto,V., Santos-Zavaleta,A., Bonavides-Martinez,C. *et al.* (2004) RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12. *Nucleic Acids Res.*, **32**, D303–D306.
23. Kunst,F., Ogasawara,N., Moszer,I., Albertini,A.M., Alloni,G., Azevedo,V., Bertero,M.G., Bessieres,P., Bolotin,A., Borchert,S. *et al.* (1997) The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature*, **390**, 249–256.
24. 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.
25. 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. USA*, **97**, 6652–6657.
26. Gough,J., Karplus,K., Hughey,R. and Chothia,C. (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.*, **313**, 903–919.
27. Madera,M., Vogel,C., Kummerfeld,S.K., Chothia,C. and Gough,J. (2004) The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Res.*, **32**, D235–D239.
28. Enright,A.J., Iliopoulos,I., Kyrpides,N.C. and Ouzounis,C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.
29. Marcotte,E.M., Pellegrini,M., Ng,H.L., Rice,D.W., Yeates,T.O. and Eisenberg,D. (1999) Detecting protein function and protein–protein interactions from genome sequences. *Science*, **285**, 751–753.
30. Fraser,C.M., Casjens,S., Huang,W.M., Sutton,G.G., Clayton,R., Lathigra,R., White,O., Ketchum,K.A., Dodson,R., Hickey,E.K. *et al.* (1997) Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature*, **390**, 580–586.
31. Oshima,K., Kakizawa,S., Nishigawa,H., Jung,H.Y., Wei,W., Suzuki,S., Arashida,R., Nakata,D., Miyata,S., Ugaki,M. *et al.* (2004) Reductive evolution suggested from the complete genome sequence of a plant-pathogenic phytoplasma. *Nature Genet.*, **36**, 27–29.
32. Sasaki,Y., Ishikawa,J., Yamashita,A., Oshima,K., Kenri,T., Furuya,K., Yoshino,C., Horino,A., Shiba,T., Sasaki,T. *et al.* (2002) The complete genomic sequence of *Mycoplasma penetrans*, an intracellular bacterial pathogen in humans. *Nucleic Acids Res.*, **30**, 5293–5300.
33. Neyrolles,O., Chambaud,I., Ferris,S., Prevost,M.C., Sasaki,T., Montagnier,L. and Blanchard,A. (1999) Phase variations of the *Mycoplasma penetrans* main surface lipoprotein increase antigenic diversity. *Infect. Immun.*, **67**, 1569–1578.
34. Riley,M. (1998) Systems for categorizing functions of gene products. *Curr. Opin. Struct. Biol.*, **8**, 388–392.
35. Serres,M.H., Goswami,S. and Riley,M. (2004) GenProtEC: an updated and improved analysis of functions of *Escherichia coli* K-12 proteins. *Nucleic Acids Res.*, **32**, D300–D302.
36. Steeghs,L., Jennings,M.P., Poolman,J.T. and van der Ley,P. (1997) Isolation and characterization of the Neisseria meningitidis lpxD-fabZ-lpxA gene cluster involved in lipid A biosynthesis. *Gene*, **190**, 263–270.
37. Polarek,J.W., Williams,G. and Epstein,W. (1992) The products of the kdpDE operon are required for expression of the Kdp ATPase of *Escherichia coli*. *J. Bacteriol.*, **174**, 2145–2151.
38. Clegg,S., Yu,F., Griffiths,L. and Cole,J.A. (2002) The roles of the polytopic membrane proteins NarK, NarU and NirC in *Escherichia coli* K-12: two nitrate and three nitrite transporters. *Mol. Microbiol.*, **44**, 143–155.
39. Sain,B. and Murray,N.E. (1980) The hsd (host specificity) genes of *E.coli* K 12. *Mol. Gen. Genet.*, **180**, 35–46.
40. Yamanaka,K., Ogura,T., Niki,H. and Hiraga,S. (1992) Identification and characterization of the smbA gene, a suppressor of the mukB null mutant of *Escherichia coli*. *J. Bacteriol.*, **174**, 7517–7526.
41. Ohnishi,M., Janosi,L., Shuda,M., Matsumoto,H., Hayashi,T., Terawaki,Y. and Kaji,A. (1999) Molecular cloning, sequencing,

purification, and characterization of *Pseudomonas aeruginosa* ribosome recycling factor. *J. Bacteriol.*, **181**, 1281–1291.

42. Gerth,U., Wipat,A., Harwood,C.R., Carter,N., Emmerson,P.T. and Hecker,M. (1996) Sequence and transcriptional analysis of clpX, a class-III heat-shock gene of *Bacillus subtilis*. *Gene*, **181**, 77–83.
43. Podkovyrov,S. and Larson,T.J. (1995) Lipid biosynthetic genes and a ribosomal protein gene are cotranscribed. *FEBS Lett.*, **368**, 429–431.
44. Green,S.M., Malik,T., Giles,I.G. and Drabble,W.T. (1996) The purB gene of *Escherichia coli* K-12 is located in an operon. *Microbiology*, **142** (Pt 11), 3219–3230.
45. Wolfe,A.J., Conley,M.P., Kramer,T.J. and Berg,H.C. (1987) Reconstitution of signaling in bacterial chemotaxis. *J. Bacteriol.*, **169**, 1878–1885.
46. Macpherson,D.F., Manning,P.A. and Morona,R. (1994) Characterization of the dTDP-rhamnose biosynthetic genes encoded in the rfb locus of *Shigella flexneri*. *Mol. Microbiol.*, **11**, 281–292.
47. Szabo,M., Bronner,D. and Whitfield,C. (1995) Relationships between rfb gene clusters required for biosynthesis of identical D-galactose-containing O antigens in *Klebsiella pneumoniae* serotype O1 and *Serratia marcescens* serotype O16. *J. Bacteriol.*, **177**, 1544–1553.
48. Nystrom,T. (1994) The glucose-starvation stimulon of *Escherichia coli*: induced and repressed synthesis of enzymes of central metabolic pathways and role of acetyl phosphate in gene expression and starvation survival. *Mol. Microbiol.*, **12**, 833–843.
49. Kakuda,H., Hosono,K., Shiroishi,K. and Ichihara,S. (1994) Identification and characterization of the ackA (acetate kinase A)-pta (phosphotransacetylase) operon and complementation analysis of acetate utilization by an ackA-pta deletion mutant of *Escherichia coli*. *J. Biochem.*, **116**, 916–922.
50. Downing,W.L., Sullivan,S.L., Gottesman,M.E. and Dennis,P.P. (1990) Sequence and transcriptional pattern of the essential *Escherichia coli* secE-nusG operon. *J. Bacteriol.*, **172**, 1621–1627.
51. Anantharaman,V. and Aravind,L. (2003) Evolutionary history, structural features and biochemical diversity of the NlpC/P60 superfamily of enzymes. *Genome Biol.*, **4**, R11.
52. Yoshimoto,T., Nagai,H., Ito,K. and Tsuru,D. (1993) Location of the 7 alpha-hydroxysteroid dehydrogenase gene (hdhA) on the physical map of the *Escherichia coli* chromosome. *J. Bacteriol.*, **175**, 5730.
53. Tyson,G.W., Chapman,J., Hugenholtz,P., Allen,E.E., Ram,R.J., Richardson,P.M., Solovyev,V.V., Rubin,E.M., Rokhsar,D.S. and Banfield,J.F. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, **428**, 37–43.
54. Venter,J.C., Remington,K., Heidelberg,J.F., Halpern,A.L., Rusch,D., Eisen,J.A., Wu,D., Paulsen,I., Nelson,K.E., Nelson,W. *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66–74.