evolution of eukaryotic chromosomes. *Proc. Natl. Acad. Sci. U. S. A.* 89, 1075–1079

11 Hill, W.G. and Robertson, A. (1966) The effect of linkage on limits to artificial selection. *Genet. Res.* 8, 269–294

12 Carvalho, A.B. and Clark, A.G. (1999) Intron size and natural selection. *Nature* 401, 344

13 Comeron, J.M. and Kreitman, M. (2000) The correlation between intron length and recombination in *Drosophila*. Dynamic equilibrium between mutational and selective forces. *Genetics* 156, 1175–1190

14 Charlesworth, B. (1996) Genome evolution – the changing sizes of genes. *Nature* 384, 315–316

15 Hurst, L.D. *et al.* (1999) Small introns tend to occur in GC-rich regions in some but not all vertebrates. *Trends Genet.* 15, 437–439

16 Petrov, D.A. *et al.* (2000) Evidence for DNA loss as a determinant of genome size. *Science* 287, 1060–1062

17 Otto, S.P. and Barton, N.H. (1997) The evolution of recombination: removing the limits to natural selection. *Genetics* 147, 879–906

18 Hey, J. (1998) Selfish genes, pleiotropy and the origin of recombination. *Genetics* 149, 2089–2097

19 Ophir, R. and Graur, D. (1997) Patterns and rates of indel evolution in processed pseudogenes from humans and murids. *Gene* 205, 191–202

20 Smit, A.F. (1999) Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* 9, 657–663

21 Duret, L. and Hurst, L.D. The elevated G and C content at exonic third sites is not evidence against neutralist models of isochore evolution. *Mol. Biol. Evol.* (in press)

22 Duret, L. *et al.* (2000) Transposons but not retrotransposons are located preferentially in regions of high recombination rate in *Caenorhabditis elegans*. *Genetics* 156, 1661–1669

L. Duret
Laboratoire de Biométrie et Biologie Évolutive, CNRS UMR 5558,
Université Claude Bernard–Lyon 1,
43 Bd du 11 Novembre 1918,
69622 Villeurbanne,
Cedex, France.
e-mail: duret@biomserv.univ-lyon1.fr

Genome Analysis

Transcription unit conservation in the three domains of life: a perspective from *Escherichia coli*

Gabriel Moreno-Hagelsieb, Victor Treviño, Ernesto Pérez-Rueda, Temple F. Smith and Julio Collado-Vides

Here we address the question of the degree to which genes within experimentally characterized operons in one organism (*Escherichia coli*) are conserved in other genomes. We found that two genes adjacent within an operon are more likely both to have an ortholog in other organisms, regardless of relative position, than genes adjacent on the same strand but in two different transcription units. They are also more likely to occur next to, or fused to, one another in other genomes. Genes frequently conserved adjacent to each other, especially among evolutionarily distant species, must be part of the same transcription unit in most of them.

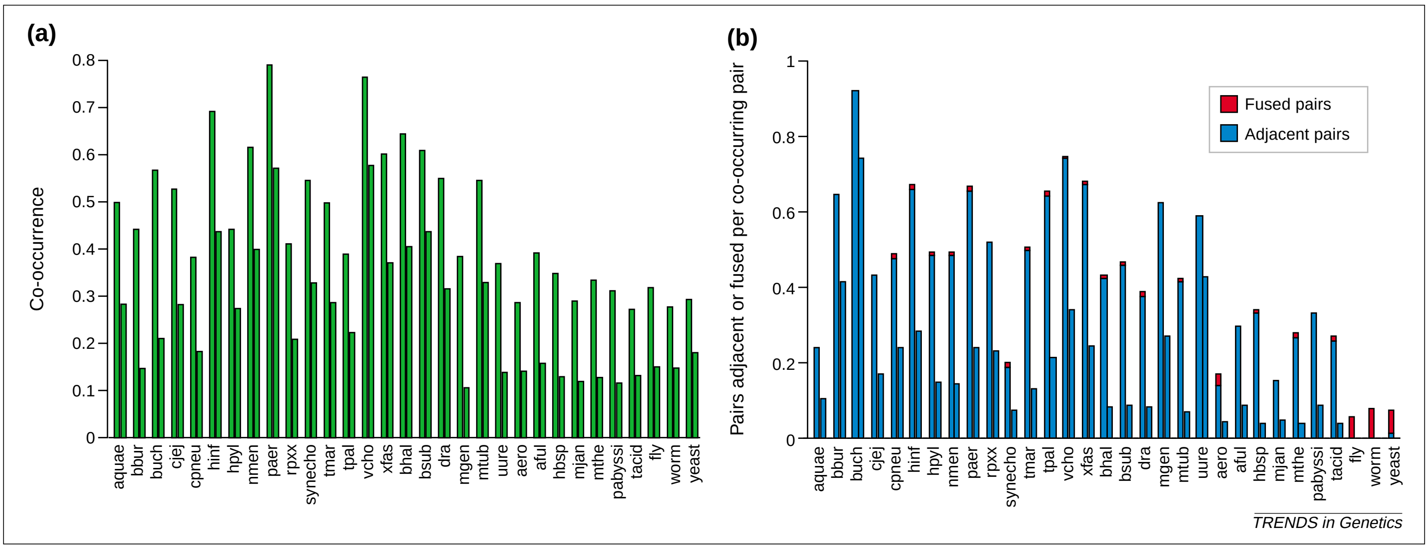
Analyses of genome organization have shown that gene order differs between genomes[1], and that such order deteriorates much faster than protein sequence identities[2]. Despite this, it has been possible to find some conserved gene clusters with protein products that either physically interact[3] or have an otherwise related function[4]. Such clusters have been related to operons implicitly in the texts and explicitly in the examples. Nonetheless, if we understand operons as a collection of adjacent genes transcribed into a single messenger RNA, or polycistronic transcription unit (TU), there is still a need to demonstrate that the conserved clusters correspond to operons.

Here, we demonstrate that genes within experimentally characterized operons in *Escherichia coli* have evident tendencies towards conservation in other genomes, and that pairs of genes showing a high conservation of vicinity, might actually be part of the same TU in most prokaryotic organisms. Our computer analyses were based on the comparison of the conservation of adjacent pairs of genes transcribed in the same direction in *E. coli*. The genes were from two collections built as described previously[5]: a collection of pairs found in operons in RegulonDB, a database compiled from the literature on regulation of transcription in *E. coli*[6] (612 pairs out of 269 operons), and, as a control, a dataset of adjacent pairs found at the boundaries of TUs (405 pairs); that is, the last gene in a TU, and the first in the next one. We call these two sets 'within-operon pairs' and 'boundary pairs', respectively.

To find orthologs of *E. coli* genes in other organisms, we ran gapped BLASTP (Ref. 7) comparisons of all the protein sequences corresponding to all the open reading frames (ORFs) of *E. coli*, against every protein sequence corresponding to the ORFs of all other genomes obtained from GenBank (Ref. 8). We used an expectation value cutoff of 0.01. We kept only those results where the alignment covered at least 50% of one of the sequences. Our putative orthologs were those genes with protein products that were overall best hits to the *E. coli* query proteins. We used this uni-directional best hits definition of orthology instead of the more common bi-directional one (i.e. the query is also the best hit when its best hit is used as query), because we observed that the data self-clean as the analyses advance. The definition also facilitates the finding of fusions and of synteny (conservation of gene order), which is another indication of orthology[2].

Co-occurrence of genes among genomes

If the proteins coded by two genes have a related function (for instance, take part in sequential steps of a pathway), they would be expected to co-occur in different genomes. Thus, it has been suggested[2] and demonstrated[9], that functional relationships of genes can be inferred if such genes have similar 'phylogenetic profiles'. Figure 1a shows that this trend characterizes adjacent genes within operons, most of which are formed from functionally related genes. If the two members of an *E. coli* pair each have an ortholog in another genome, regardless of the relative positions of these orthologs within that genome, we call them an ortholog pair, or a co-occurring pair. If only



TRENDS in Genetics

Fig. 1. Conservation of pairs of adjacent genes known to be in operons in *Escherichia coli* compared with that of genes at the borders of transcription units. (a) Pair co-occurrence. We measure co-occurrence as the number of pairs of adjacent genes where both genes have an ortholog in the genome of interest, normalized against the sum of such pairs and pairs where only one of the genes in the pair has an ortholog (orphan orthologs). (b) Pairs conserved adjacent to each other in the same strand or fused together. The first column in each organism represents the gene pairs with orthologous pairs contained within *E. coli* operons, and the second represents gene pairs with orthologous pairs occurring at the borders of *E. coli* transcription units. Note that the column representing pairs in operons is always higher and that fusions occur only among conserved genes corresponding to operons. The labels mostly correspond to the file names at GenBank: aquae, *Aquifex aeolicus*; bbur, *Borrelia burgdorferi*; buch, *Buchnera* sp. APS; cjej, *Campylobacter jejuni*; cpneu, *Chlamydia pneumoniae*; hinf, *Haemophilus influenzae*; hpyl, *Helicobacter pylori* 26695; nmen, *Neisseria meningitidis* MC58; paer, *Pseudomonas aeruginosa*; rpxx, *Rickettsia prowazekii*; synecho, *Synechocystis* PCC6803; tmar, *Thermotoga maritima*; tpal, *Treponema pallidum*; vcho, *Vibrio cholerae*; xfas, *Xylella fastidiosa*; bhal, *Bacillus halodurans*; bsub, *Bacillus subtilis*; dra, *Deinococcus radiodurans*; mgen, *Mycoplasma genitalium*; mtub, *Mycobacterium tuberculosis*; uure, *Ureaplasma urealyticum*; aero, *Aeropyrum pernix*; aful, *Archaeoglobus fulgidus*; hbsp, *Halobacterium* sp. NRC-1; mjan, *Methanococcus jannaschii*; mthe, *Methanobacterium thermoautotrophicum*; pabyssi, *Pyrococcus abyssi*; tacid, *Thermoplasma acidophilum*; fly, *Drosophila melanogaster*; worm, *Caenorhabditis elegans*; yeast, *Saccharomyces cerevisiae*.

one of the genes in a pair has an ortholog in another genome, the ortholog is called an ortholog orphan. We measure co-occurrence as the number of ortholog pairs, normalized against the sum of ortholog pairs plus ortholog orphans. Co-occurrence of within-operon pairs is higher among genomes than co-occurrence of boundary pairs. For instance, in *Aquifex aeolicus*, there are 197 ortholog pairs and 196 orphans corresponding to *E. coli* within-operon pairs. This makes 197/(197+196) or 0.5 of co-occurrence, which is higher than the 0.28 of co-occurrence corresponding to the boundary pairs dataset.

Conservation of neighborhood

Once we found all the corresponding ortholog pairs, we verified the relative locations of their gene members. We considered that two genes were conserved as neighbors when the corresponding orthologs were found adjacent to each other on the same strand or were fused into a single gene, in the other genome. Conservation of neighborhood[3,4] and appearance of fusions[10–12] are also useful hints in detecting functionally related genes. Fusions were only found among orthologs to *E. coli* within-operon pairs, and conservation of vicinity is clearly higher among these pairs than among boundary pairs (Fig. 1b). Fusions accounted for just a few hits within prokaryotic genomes (zero or one in most species and a maximum of three in *Deinococcus radiodurans*, *Haemophilus influenzae*, *Pseudomonas aeruginosa*, *Synechocystis* sp. and *Vibrio cholerae*), whereas in Eukarya, all pairs found together, except one in yeast, form fused genes. In *A. aeolicus*, there are 47 pairs of genes adjacent to each other out of 197 ortholog pairs to *E. coli* within-operon pairs. Thus, about 24% of the pairs found are kept together, whereas only 10% of the co-occurring boundary pairs are neighbors in *A. aeolicus*. A similar trend is observed in all other genomes.

Conserved neighborhood and organization of TUs

We have previously demonstrated that *E. coli* within-operon pairs can be distinguished from boundary pairs using a method derived from their distinctive frequency distribution of intergenic distances[5]. If within operons pairs can be distinguished in the same way in other organisms, we would expect pairs of genes conserved as neighbors to display similar distributions to that of the collection of pairs to which they correspond. We plotted the frequency distributions of the intergenic distances of genes conserved adjacent to each other among genomes, against those of adjacent genes within operons in *E. coli* and of genes at TU boundaries in *E. coli* (Fig. 2). As expected, the frequency distribution of intergenic distances of ortholog pairs corresponding to within-operon pairs is very similar to that of genes within operons in *E. coli* (Fig. 2a). Nevertheless, ortholog boundary pairs display an intermediate distribution (Fig. 2b). This result indicates that conserved boundary pairs are a mixed population, as further analyses confirm. Most of the pairs found at TU boundaries in *E. coli* are conserved as neighbors in no more than two other genomes. However, some pairs are conserved as neighbors in more than five and up to 23 other genomes; these genes have intergenic distances typical of genes within operons. This implies that such genes are in operons in many other organisms and might have recently split into another TU in *E. coli*.

We were able to provide further proof of this assumption. A collection of 100 operons of *Bacillus subtilis* compiled from the literature is available[13]. We used it to build a dataset of within-operon pairs (310 pairs). We also used this collection of operons to find TU boundaries by comparison with the genome sequence and annotation for this organism[14], and built a dataset of boundary pairs (123 pairs). Because we already found which pairs in the *E. coli* datasets are

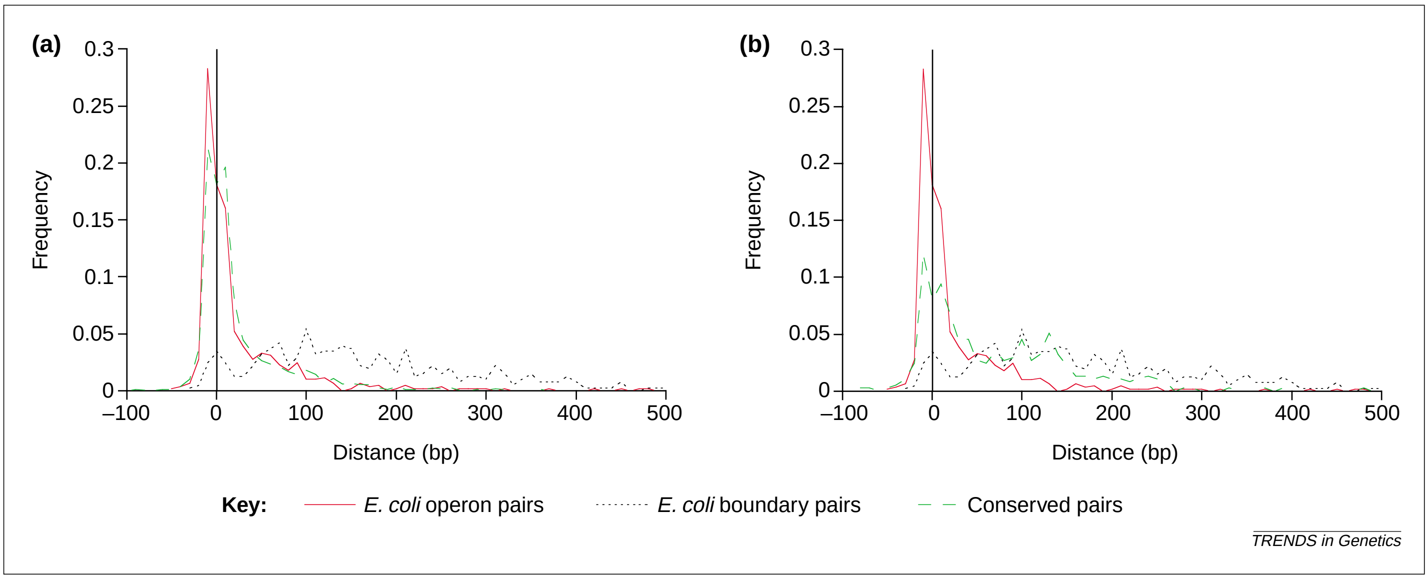


Fig. 2. Comparison of the frequency distribution of intergenic distances between pairs of genes in operons, pairs at the borders of transcription units (TU boundaries), and pairs conserving their neighborhood in other organisms. (a) Conserved pairs of genes within operons. (b) Conserved pairs of genes at TU boundaries. The mixed population in conserved TU boundaries is evidenced by the small peak at the same place as the peak in the *E. coli* operon curve and the coincidence of the rest of the curve with that of *E. coli* TU boundaries.

conserved as neighbors in *B. subtilis*, we found the intersections between such conserved pairs and the datasets built in *B. subtilis*. Fifty-nine out of 62 (about 95%) pairs of genes found in operons in *E. coli*, are also in operons in *B. subtilis* (the remaining three are boundary pairs in *B. subtilis*). Among those found at TU boundaries in *E. coli*, three out of four are in operons in *B. subtilis*, and are conserved as neighbors in at least 12 other genomes. Other conserved boundary pairs should represent pairs kept together by chance: inheritance from a common ancestor, especially between closely related species and, much less probably, coincidental rearrangements with no further biological implications. Other causes might be involved, but our results strongly suggest that organization into operons is the main reason for conservation of adjacency between evolutionarily distant species.

Concluding remarks

The inference of functional relationships from genomic context (phylogenetic profiles, conservation of vicinity, gene fusions) has had considerable attention lately (e.g. Refs 15–18). The comparisons shown in the figures here provide a new perspective of the differences expected from contrasting populations whose relationships are known from experiment (genes within operons, mostly known to have a functional relationship, against genes at TU boundaries, which do not necessarily have functional relationships). We have also shown that orthologs to within-operon pairs have a similar trend towards keeping short intergenic distances. This gives the first glimpse that TU predictions, based on intergenic distances, as implemented in *E. coli*[5], can be applied to other prokaryotic genomes.

Acknowledgements

This work was supported by grant number 0028 from Consejo Nacional de Ciencia y Tecnología to J.C-V., and by US Dept of Energy grant number DE-FG02-98ER62558. We acknowledge useful comments from an anonymous referee and technical support from V. del Moral.

References

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

2 Huynen, M.A. and Bork, P. (1998) Measuring genome evolution. *Proc. Natl. Acad. Sci. U. S. A.* 95, 5849–5856

3 Dandekar, T. *et al.* (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.* 23, 324–328

4 Overbeek, R. *et al.* (1999) The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. U. S. A.* 96, 2896–2901

5 Salgado, H. *et al.* (2000) Operons in *Escherichia coli*: genomic analyses and predictions. *Proc. Natl. Acad. Sci. U. S. A.* 97, 6652–6657

6 Salgado, H. *et al.* (2000) RegulonDB (version 3.0): transcriptional regulation and operon organization in *Escherichia coli* K-12. *Nucleic Acids Res.* 28, 65–67

7 Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402

8 Benson, D.A. *et al.* (2000) GenBank. *Nucleic Acids Res.* 28, 15–18

9 Pellegrini, M. *et al.* (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U. S. A.* 96, 4285–4288

10 Das, S. *et al.* (1997) Biology's new Rosetta stone. *Nature* 385, 29–30

11 Enright, A.J. *et al.* (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature* 402, 86–90

12 Marcotte, E.M. *et al.* (1999) Detecting protein function and protein-protein interactions from genome sequences. *Science* 285, 751–753

13 Itoh, T. *et al.* (1999) Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. *Mol. Biol. Evol.* 16, 332–346

14 Kunst, F. *et al.* (1997) The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* 390, 249–256

15 Huynen, M. *et al.* (2000) Exploitation of gene context. *Curr. Opin. Struct. Biol.* 10, 366–370

16 Huynen, M. *et al.* (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res.* 10, 1204–1210

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

18 Eisenberg, D. *et al.* (2000) Protein function in the post-genomic era. *Nature* 405, 823–826

G. Moreno-Hagelsieb*
V. Treviño
E. Pérez-Rueda
J. Collado-Vides
Laboratory of Computational Biology, CIFN, UNAM, A.P. 565-A, Cuernavaca, Morelos 62100, Mexico.
*e-mail: moreno@cifn.unam.mx

T.F. Smith
Biomolecular Engineering Research Center, Boston University, 36 Cummington St, Boston, MA 02115, USA.