nature
biotechnology

Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs

Jan O Korbel, Lars J Jensen, Christian von Mering & Peer Bork

Several widely used methods for predicting functional associations between proteins are based on the systematic analysis of genomic context. Efforts are ongoing to improve these methods and to search for novel aspects in genomes that could be exploited for function prediction. Here, we use gene expression data to demonstrate two functional implications of genome organization: first, chromosomal proximity indicates gene coregulation in prokaryotes independent of relative gene orientation; and second, adjacent bidirectionally transcribed genes (that is,'divergently' organized coding regions) with conserved gene orientation are strongly coregulated. We further demonstrate that such bidirectionally transcribed gene pairs are functionally associated and derive from this a novel genomic context method that reliably predicts links between >2,500 pairs of genes in ~100 species. Around 650 of these functional associations are supported by other genomic context methods. In most instances, one gene encodes a transcriptional regulator, and the other a nonregulatory protein. In-depth analysis in *Escherichia coli* shows that the vast majority of these regulators both control transcription of the divergently transcribed target gene/operon and auto-regulate their own biosynthesis. The method thus enables the prediction of target processes and regulatory features for several hundred transcriptional regulators.

Several bioinformatics methods have recently been developed that exploit the genomic context of genes to predict functions for the encoded proteins. These approaches analyze gene fusion[1,2], gene neighborhood (that is, the assembly of genes in putative operons[3,4]), and the co-occurrence of genes across genomes[5] to predict 'functional associations' for a given protein—for example, physical interaction partners or members of the same biochemical pathway (see **Box 1**). Numerous modifications and combinations of the original approaches have been published (*e.g.*, see refs. 6–12 and references therein). In addition to genomic context methods that require no information other than sets of open reading frames from completely sequenced genomes, other data types such as protein structures, DNA-regulatory elements, or experimental results from DNA microarrays and physical interaction screens have been used recently for protein interaction prediction (for selected reviews, see refs. 13–21; see also **Box 1**).

Despite the number of approaches—and regardless of the power of homology-based function prediction[22–24]—a considerable amount of the proteins present in databases remain uncharacterized. For instance, the well-curated COG (clusters of orthologous groups) database[25] contains around 10,000 prokaryotic proteins in almost 1,000 orthologous groups that still lack functional assignments.

Several groups, including ours, have been searching for additional functional signals in genomic context that can be exploited to extend or complement previous methods for protein function prediction. For example, anticorrelating occurrences of genes across genomes have been used to identify instances where a known enzyme may have been displaced by a novel, functionally equivalent protein[26].

Here, we propose a new approach that exploits the conservation of divergently (bidirectionally) transcribed gene pairs. The method is complementary to the existing gene neighborhood method, which focuses on clusters of genes found in putative operons (that is, genes transcribed in a common orientation, or codirectionally). It is generally accepted that genes within an operon are almost always functionally associated[27,28]. However, a few experimental case studies report functional relationships also for adjacent, divergently transcribed genes[29–31], which are presumably linked by overlapping promoter regions[29,32]. Bidirectional gene organization may be advantageous, as it provides a means of transcriptional coregulation distinct from operons. Moreover, although divergently transcribed genes were described in prokaryotes more than 30 years ago (see ref. 31 and references therein), bidirectional gene organization has never been studied systematically in the light of protein function prediction.

Conservation of gene orientation in prokaryotes

Generally, evolutionary conservation is a good indicator of functional relevance, as it allows biologically relevant signals to be discerned from background noise[3,4,33,34]. In prokaryotes, codirectional gene neighbors are well conserved because of the prevalent operons of functionally related genes. However, as has been noted previously[35,36], the

European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany, and Max Delbrück Center for Molecular Medicine, 13092 Berlin-Buch, Germany. Correspondence should be addressed to P.B. (bork@embl-heidelberg.de).

Published online 30 June 2004; doi:10.1038/nbt988

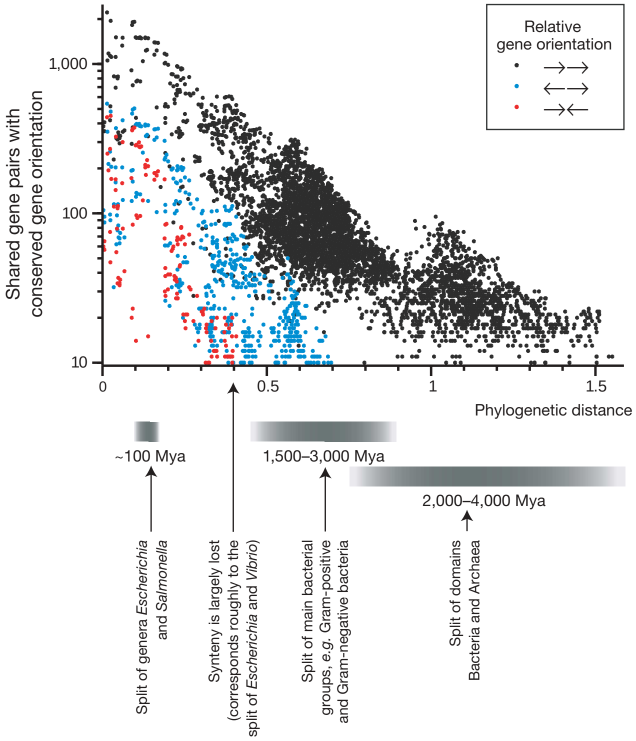


Figure 1 Evolutionary conservation of the orientation of gene neighbors across prokaryotic lineages. We show an updated version of figures published when 32 prokaryotic genomes had already been sequenced[35,36], providing a first indication of the intriguing conservation of divergently transcribed gene pairs (DT-pairs). Here, we compare 101 prokaryotic genomes in a pairwise fashion. Each dot corresponds to a pair of species; we plot numbers of adjacent codirectionally (→ →), divergently (← →) and convergently (→ ←) transcribed gene neighbors conserved across species versus the phylogenetic distances of the species. Phylogenetic distances were derived from genomic gene content[50,51] (using SHOT[51] version 2 with default parameters). Evolutionary distances for several major lineage splits are indicated together with approximate divergence times[52–54]. Although codirectional pairs are most frequently conserved over major Eubacterial (or Archaeal) lineages, DT-pairs are also maintained in several instances. In contrast, convergently transcribed gene pairs are rapidly lost in evolution; they disappear almost entirely at a divergence time roughly corresponding to the split of *Vibrio* and *Escherichia*. Mya, million years ago.

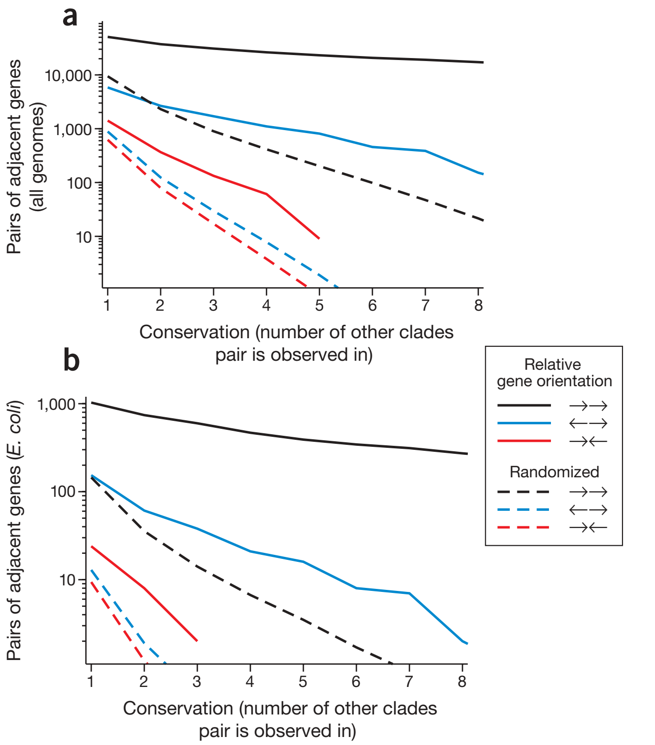


Figure 2 Conserved organization of DT-pairs. To assess the functional relevance of the three types of organization of neighboring genes, we evaluated the evolutionary conservation of relative gene orientation across prokaryotic genomes. Although codirectional pairs (→ →) are most widely conserved, pairs of divergently (← →) transcribed genes are more widely conserved than convergently (→ ←) transcribed gene pairs. Even the last type show some enrichment over random, although this can be most likely explained by synteny due to common ancestry, rather than by selection. Randomized counts were obtained by shuffling gene identities, keeping intergenic distances and numbers of divergently, convergently and codirectionally transcribed gene pairs unchanged. Randomization was repeated 4,000 times. (**a**) Pairs of adjacent genes in any of the 101 completely sequenced prokaryotic genomes for which the relative gene orientation has been conserved across evolutionary clades. Cumulative counts are shown for varying degrees of conservation. (**b**) Cumulative counts of conserved gene pairs observed in *E. coli*.

organization of divergently transcribed gene pairs (DT-pairs) is also widely conserved among prokaryotes, whereas convergently transcribed pairs are rapidly lost in evolution (**Fig. 1**). As shown in **Figure 1**, DT-pairs have often been maintained over several of the main Eubacterial (or Archaeal) branches (*e.g.*, they are retained in both Gram-positive and Gram-negative bacteria).

To systematically assess the extent and significance of DT-pair conservation, we evaluated the presence of gene pairs across the 101 completely sequenced prokaryotic genomes currently included in the STRING (search tool for the retrieval of interacting genes/proteins) database[12]. Closely related species were grouped into 47 evolutionary clades[12], within which repeated observations were considered as no more informative than single observations. We define a conserved gene pair as a pair of neighboring protein-coding genes having corresponding adjacent orthologs with the same gene orientation in a genome from another clade. Correspondingly, more widely conserved pairs have orthologs in more than one independent clade. Groups of orthologous proteins were obtained from the STRING[12] server. The orthologous groups include clusters originally obtained from the COG database[25], which were subsequently expanded and extended to accommodate more recently sequenced species[12]. (A more detailed explanation of methods is available in the **Supplementary Notes** online.)

More than 5,000 DT-pairs are conserved across evolutionarily distant clades, much more than would be expected at random (**Fig. 2a**). Depending on the level of conservation, the enrichment of DT-pairs over convergently transcribed pairs is more than 100-fold. This translates into a large number of conserved DT-pairs in any given species, most of which are probably functionally relevant.

To best characterize DT-pairs and their functional implications, we turned to the most studied prokaryotic genome, *Escherichia coli* K12, as a reference: 26% of all *E. coli* genes are part of DT-pairs, of which more than a quarter are conserved in a distant clade (**Fig. 2b**). The latter fraction is expected to increase as more genomes are sequenced. Currently, there are 6.5 times more conserved DT-pairs than conserved convergently transcribed pairs in *E. coli*, and 12 times more than expected at random. Conserved bidirectional gene orientation is

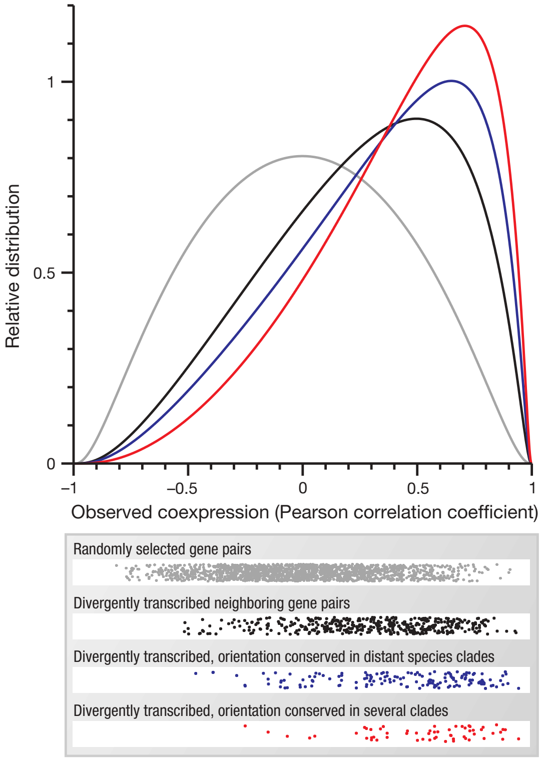


Figure 3 Correlated gene expression of adjacent divergently transcribed *E. coli* genes. Expression data from 95 previously published DNA microarrays available from SMD[37] (downloaded on 2 September 2003) were mapped onto the *E. coli* genome. For each array the cyanine (Cy)3 and Cy5 channels were individually normalized with the Qspline method, and spatial biases corrected using a Gaussian smoother[55]. The distribution of Pearson correlation coefficients is shown for 10,000 pairs of genes randomly selected from the genome, for 572 divergently transcribed neighboring gene pairs (DT-pairs), for 148 DT-pairs with evolutionarily conserved gene organization and for 58 DT-pairs conserved across several independent clades. The dot clouds show the expression correlation for individual gene pairs for which expression data were available. Dots were randomly positioned in *y*-axis direction to reduce the overlap. Above, probability density curves estimated from the dot clouds using Gaussian kernel density estimation are shown.

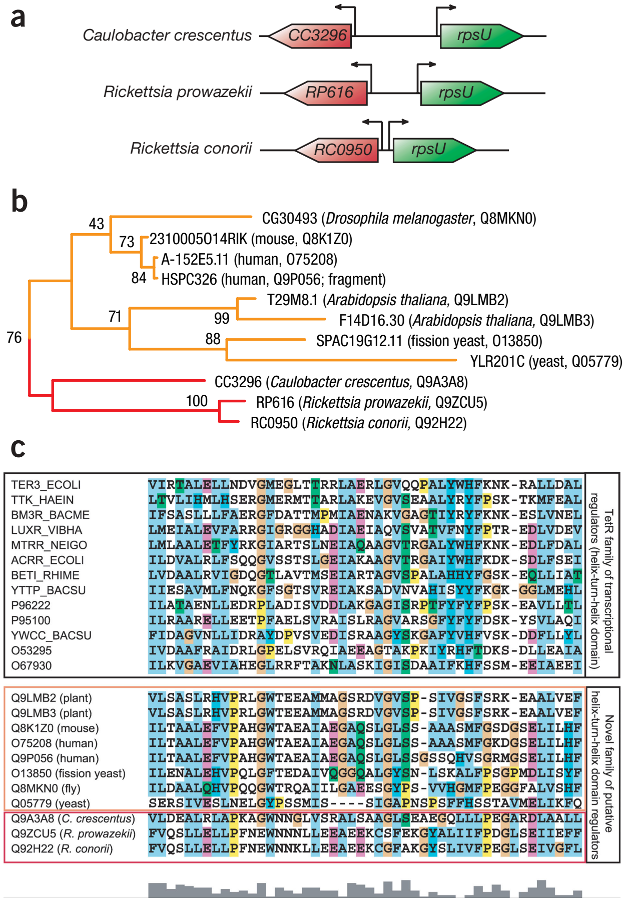

a
Caulobacter crescentus
CC3296
rpsU
Rickettsia prowazekii
RP616
rpsU
Rickettsia conorii
RC0950
rpsU
b
43
73
84
76
71
99
88
100
CG30493 (Drosophila melanogaster, Q8MKN0)
2310005014RIK (mouse, Q8K1Z0)
A-152E5.11 (human, 075208)
HSPC326 (human, Q9P056; fragment)
T29M8.1 (Arabidopsis thaliana, Q9LMB2)
F14D16.30 (Arabidopsis thaliana, Q9LMB3)
SPAC19G12.11 (fission yeast, 013850)
YLR201C (yeast, Q05779)
CC3296 (Caulobacter crescentus, Q9A3A8)
RP616 (Rickettsia prowazekii, Q9ZCU5)
RC0950 (Rickettsia conorii, Q92H22)
c
TER3_ECOLI VIRTALELLNDVGMEGLTTRRLAERLGVQQPALYWHFKNK-RALLDAL
TTK_HAEIN LTVLIHMLHSERGMERMTTARLAKEVGVSEAALYRYFPSK-TKMFEAL
BM3R_BACME IFSASLLLFAERGFDATTMPMIAENAKVGAGTIYRYFKNK-ESLVNEL
LUXR_VIBHA LMEIALEVFARRGIGRGGHADIAEIAQVSVATVFNYFPTR-EDLVDEV
MTRR_NEIGO LMLAALETFYRKGIARTSLNEIAQAAGVTRGALYWHFKNK-EDLFDAL
ACRR_ECOLI ILDVALRLFSQQGVSSTSLGEIAKAAGVTRGAIYWHFKDK-SDLFSEI
BETI_RHIME LVDAALRVIGDQGTLAVTMSEIARTAGVSPALAHHYFGSK-EQLLIAT
YTTP_BACSU IIESAVMLFNQKGFSGTSVREIAKSADVNVAHISYYFKGK-GGLMEHL
P96222 ILATAENLLEDRPLADISVDDLAKGAGISRPTFYFYFPSK-EAVLLTL
P95100 ILRAARELLEETPFAELSVRAISLRAGVARSGFYFYFDSK-YSVLAQI
YWCC_BACSU FIDAGVNLLIDRAYDPVSVEDISRAAGYSKGAFYVHFVSK-DDFLLYL
O53295 IVDAAFRAIDRLGPELSVRQIAEEAGTAKPKIYRHFTDKS-DLLEAIA
O67930 ILKVGAEVIAHEGLRRFTAKNLASKIGISDAAIFKHFSSM-EEIAEEI
TetR family of transcriptional regulators (helix-turn-helix domain)
Q9LMB2 (plant) VLSASLRHVPRLGWTEEAMMAGSRDVGVSP-SIVGSFSRK-EAALVEF
Q9LMB3 (plant) VLSASLRHVPRLGWTEEAMMAGSRDVGVSP-SIVGSFSRK-EAALVEF
Q8K1Z0 (mouse) ILTAALEFVPAHGWTAEAIAEGAQSLGLSS-AAASMFGSDGSELILHF
O75208 (human) ILTAALEFVPAHGWTAEAIAEGAQSLGLSS-AAASMFGKDGSELILHF
Q9P056 (human) ILTAALEFVPAHGWTAEAIAEGAQSLGLSSGSSQHVSGRMGSELILHF
O13850 (fission yeast) ILENALEHVPQLGFTEDAIVQGGQALGYSN-LSKALFPSGPMDLISYF
Q8MKN0 (fly) ILDAALQHVPQQGWTRQAIILGAEESGYPS-VVHGMFPEGGFALVSHF
Q05779 (yeast) SERSIVESLNELGYPSSMIS----SIGAPNSPSFFHSSTAVMELIKFQ
Q9A3A8 (C. crescentus) VLDEALRLAPKAGWNNGLVSRALSAAGLSEAEGQLLLPEGARDLAALL
Q9ZCU5 (R. prowazekii) FVQSLLELLPFNEWNNNLLEEAEEKCSFEKGYALIIFPDGLSEIIEFF
Q92H22 (R. conorii) FVQSLLELLPFNEWNNKLLEEAEEKCGFAKGYSLIVFPEGLSEIVGFL
Novel family of putative helix-turn-helix domain regulators

Figure 4 Combining homology and context information for function prediction. The enrichment of genes encoding transcriptional regulators within conserved DT-pairs suggests that novel regulators controlling their adjacent target genes may be discovered. By applying profile-based homology searches, we found evidence for novel regulatory families—for example, we predict that members of the orthologous group ‘KOG2969’ (obtained from STRING[12]) form a novel regulatory family involved in regulation of bacterial ribosomal protein S2 (*rpsU*), as well as its orthologs in eukaryotic mitochondria (see main text). (**a**) Genes encoding prokaryotic members of KOG2969 (red arrow symbols) are divergently oriented neighbors of *rpsU* (green arrow symbols). (**b**) Unrooted phylogenetic tree of KOG2969 constructed using MRBAYES[56] (see **Supplementary Notes** online). (**c**) CLUSTALW[57] sequence alignment of PFAM (protein families) database[23] domain ‘PF00440’ representing the TetR family, and the corresponding homologous region of the members of KOG2969 (amino acid sequences surrounded by colored boxed correspond to eukaryotic members (orange box), or a-proteobacterial members (red box) of KOG2969).

thus a frequently occurring phenomenon, involving a considerable fraction of prokaryotic genes.

Conserved organization of DT-pairs predicts gene coregulation

The strong evolutionary conservation of DT-pairs implies biological relevance—a plausible scenario is that, in analogy to operons, coregulation is the driving force maintaining the bidirectional orientation. We tested this hypothesis in *E. coli* by analyzing gene expression data from 95 previously published DNA microarray experiments[37], searching systematically for coexpression of neighboring genes across a variety of experimental conditions. Unexpectedly, we found a general signal of coexpression of neighbors independent of their orientation, extending over large distances (see **Box 2**).

Nevertheless, the coexpression of conserved DT-pairs rises significantly above this general background, strongly suggesting that coregulation is an evolutionary constraint maintaining their relative orientation (see **Fig. 3**, **Supplementary Notes** online; similar results were obtained when limiting the analysis to replicated microarray experiments, or to experiments from a single laboratory). Globally, conserved *E. coli* DT-pairs are more strongly coexpressed than nonconserved pairs or randomly chosen genes (both significant at the 0.001 level according to Kolmogorov-Smirnov tests). In contrast, for convergently transcribed gene pairs, there is no measurable difference in expression correlation between nonconserved pairs and the few conserved pairs (**Supplementary Notes** online), as should be expected if those pairs are only conserved by chance. A large fraction (87%) of the *E. coli* DT-pairs conserved in distant clades is positively correlated in expression—for 37%, the Pearson correlation coefficient exceeds 0.6, a threshold previously used to indicate functional association in array experiments[38]. In general, more widely conserved DT-pairs show significantly stronger expression correlation; pairs conserved across

Box 1 Comparison of genomic context approaches

Comparative analysis of genomes is the common denominator of all genomic context methods. Soon after the sequencing of the first prokaryotic genomes, the potential of exploiting the genomic context of genes for evolutionary analysis and function prediction was realized (*e.g.*, refs. 58,59). The first generation of resulting protein interaction prediction methods (reviewed in refs. 60–63) is summarized in the upper panel of **Figure 5** (the lower panel shows coverage and overlap of the methods at an accuracy level of 40%). These techniques predict general functional associations for a given protein (*e.g.*, with which other protein it interacts, or in which cellular pathway/process it is involved).

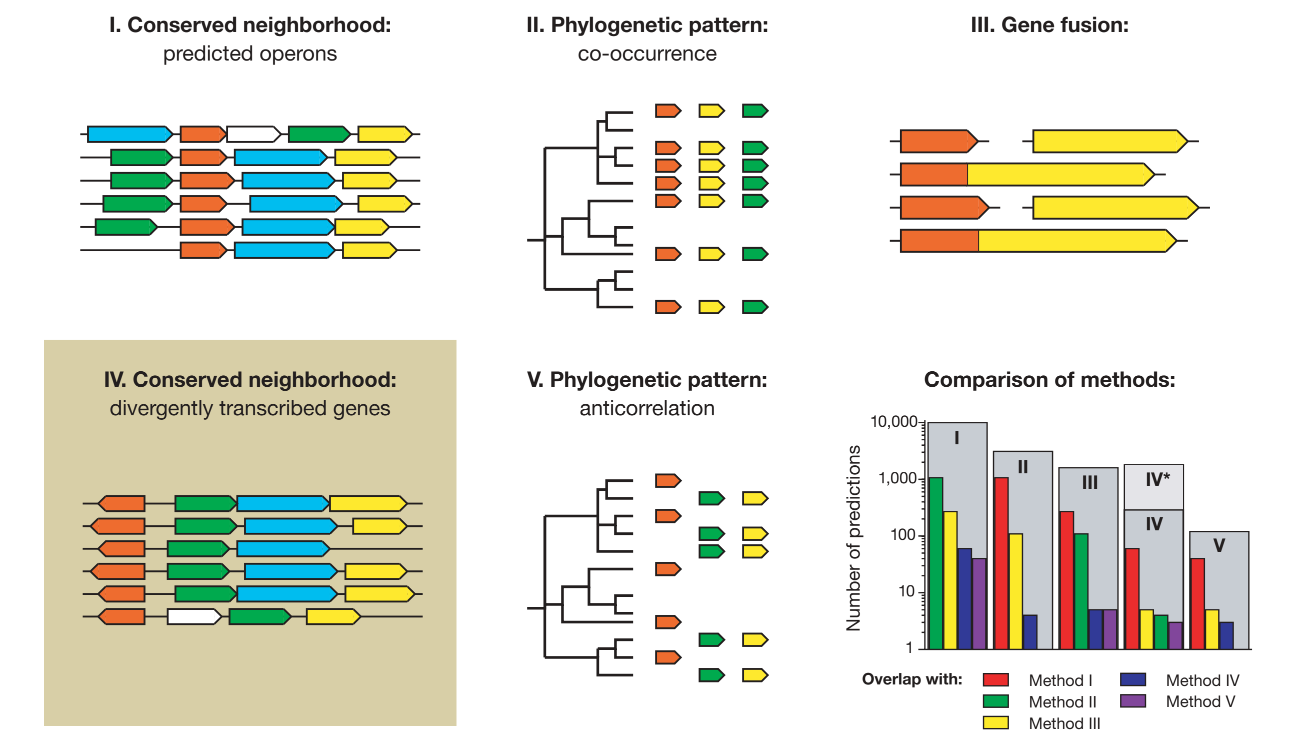


Figure 5 Different genomic context methods and their relative coverage. Method I analyzes the local gene neighborhood (evaluating gene presence in conserved, putative operons[3,4]); Method II is based on gene co-occurrence (presence and absence) across genomes[5]; Method III identifies gene fusions[1,2]; Method IV analyzes conserved DT-pairs (a variant thereof Method IV*, links all pairs of genes in conserved bidirectionally transcribed operons–excluding associations predicted by Method I); and Method V detects functionally equivalent genes by analyzing anticorrelating gene occurrences[26].

Many implementations with different flavors exist by now (*e.g.*, see refs. 6–12) and conceptually distinct methods have been developed that explore similar evolutionary signals (*e.g.*, phylogenetic trees versus co-occurrence profiles, *e.g.*, ref. 64). Nowadays, gene context approaches are routinely used to predict protein functions in prokaryotes; they have identified numerous novel functions that were later experimentally verified. These include physical interactions between proteins in signal transduction[65,66] and DNA repair[67], novel members of metabolic pathways (*e.g.*, ref. 68) and general cellular processes like iron-sulfur protein maturation[69,70].

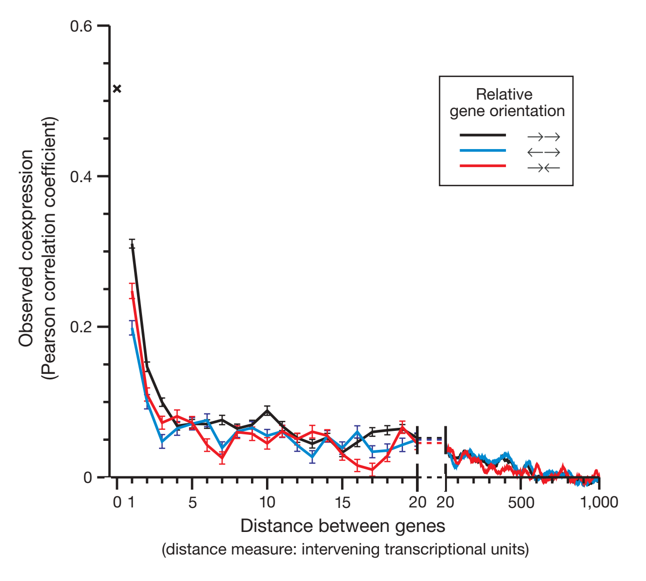
Recently, a new generation of genomic context methods has been put forth. Although these reveal fewer functional associations than the former, they allow more detailed predictions of protein function. As described in this article, the conserved organization of adjacent divergently transcribed genes (Method IV) predicts functional associations between the genes, allowing precise statements about the type of association, namely, transcriptional regulation.

Method V is based on the observation that anticorrelating occurrences of genes across genomes may indicate displacement of functionally equivalent genes[61,71]. Recent implementation of this technique led to the discovery of several analogous enzyme displacements in the thiamin biosynthesis pathway[26]. Thus, not merely involvement in a particular cellular process, but precise functional roles can be predicted. Preliminary numbers from a global analysis (see **Fig. 5**) indicate that several novel gene displacements can be predicted in a global study (data not shown).

Accuracy and coverage of genomic context methods depend on the species and functional system studied, but often also on implementation details. For instance, in the case of Method IV, conserved operon structures frequently extend from at least one side of a conserved *DT-pair*, as indicated in **Figure 5**. As conserved operons strongly indicate functional association[3,4], it is possible to predict links between all pairs of genes in conserved bidirectionally transcribed operon structures (not including those already found by Method I). Applying this strategy (Method IV*) improves the coverage but lowers the prediction accuracy slightly compared with Method IV, in which only direct neighbors are considered (see **Supplementary Notes** online for more details).

Box 2 Genomic vicinity and gene coexpression in prokaryotes

It is well known that the expression of a prokaryotic gene is not independent of its position in the genome. Because of the existence of operons[72]—clusters of functionally related genes transcribed as a single mRNA—adjacent genes tend to be coexpressed if transcribed in the same direction[73]. This effect was thought to be relatively short-range; the average operon length in *E. coli* is approximately three genes[74]. Prokaryotic gene expression correlates also with the distance to the origin of replication, as DNA close to the origin may be present in higher copy number within the cell[75]. This correlation, although weaker, should act over considerably longer distances than the influence of operons.



Recently, a large number of microarray expression data sets became available for prokaryotes[37], allowing assessment of the effects mentioned above by correlating gene expression with position and orientation of genes. Surprisingly, when mapping microarray expression data onto the genome sequence of *E. coli* K12 (data were treated as in **Fig. 3**), we find striking correlations in expression between nearby loci (see **Fig. 6**). The correlation is largely independent of relative gene orientation and can thus not be fully explained by the existence of operons. Also, even genes that are more than 100 kb apart show significant coexpression in excess of what would be expected from the influence of the replication origin (**Fig. 6** and **Supplementary Notes** online). At short distances, some of the observed correlation in expression may be due to 'transcriptional read-through'[73], that is, deferred termination of transcription (cDNA-based microarrays cannot distinguish sense from antisense transcription). However, this should primarily affect convergently transcribed genes, having extremely short intergenic distances. In contrast, divergently transcribed genes may be affected by 'read-through' only if transcription initiates outside the pair and proceeds erroneously through both genes.

Together with the recent observation of unexpectedly large regions of coexpressed genes in the filamentous cyanobacterium *Anabaena* sp.[76], our results indicate the existence of a higher order gene regulatory mechanism in many if not all prokaryotes—possibly involving changes in DNA supercoiling[77] or DNA compaction by proteins, such as Fis, H-NS or HU[78].

Figure 6 Genomic vicinity indicates gene coexpression. Correlations in DNA microarray expression were determined between all possible pairs of genes in *E. coli*, and average correlation in expression plotted versus distance between genes. Distances are given in 'transcriptional units' (tu), that is, we counted the number of operons or independently regulated genes positioned between the genes in question (neighboring codirectional genes with a distance of ≤40 bp were predicted to reside in the same *E. coli* operon[28]). Adjacent genes that are not part of the same predicted operon have a distance of tu = 1. Expression correlations for three types of relative orientation are shown: genes transcribed codirectionally (→ →), divergently (← →) and convergently (→ ←). '×' indicates the average expression correlation of codirectional genes predicted to reside in the same operon (*i.e.*, where tu = 0). 1 tu corresponds to ~1.7 kb.

three or more clades have an expression correlation comparable to that of genes in predicted *E. coli* operons (see **Supplementary Notes** online for more details).

Excess of self-regulatory transcription factors in DT-pairs

Conserved DT-pairs can provide more specific functional relationships than just predicting coregulation. Within the set of conserved DT-pairs, we observed a strong enrichment of pairs in which one gene encodes a transcriptional regulator (R), and the other gene encodes any other class of protein (X). This suggests that precise regulatory interactions may be predicted from conserved DT-pairs. Considering all prokaryotic genomes, over 71% of the DT-pairs conserved across at least four clades are classified as RX (1,133 in total), a fraction rising to 100% for more widely conserved pairs. On the basis of shuffled genomes, we conservatively estimate an RX enrichment of 1.8- to 6.5-fold, depending on pair conservation (**Supplementary Notes** online).

In contrast, pairs classified as regulator-regulator pairs (RR) or pairs where neither gene is a known transcriptional regulator (XX) are strongly underrepresented and entirely absent among widely conserved pairs. Notably, a considerable fraction of characterized proteins encoded by DT-pairs classified as XX may in fact act as post-transcriptional regulators: we found evidence for at least five such cases among the 18 most widely conserved *E. coli* DT-pairs with XX classification (**Supplementary Notes** online). Furthermore, for several poorly characterized ('hypothetical') proteins classified as X, we found homology to known transcriptional regulators (see *e.g.*, **Fig. 4**).

To test whether transcriptional regulators encoded in conserved DT-pairs tend to directly regulate the adjacent divergently transcribed gene, we analyzed all 135 regulators of *E. coli* having at least one known, annotated target gene[39–41]. Of the subset of 22 of these regulators positioned within evolutionarily conserved DT-pairs, at least 13 have been described to regulate the respective adjacent gene (and recent evidence supports one additional instance; see **Supplementary Notes** online). Thus, nearly two-thirds of these regulators have already been implicated in controlling the divergently transcribed gene—the actual number is probably higher, as almost all of the remaining adjacent genes have no annotated regulator[39–41] and genes may be controlled by more than one regulatory protein[39].

Interestingly, evolutionarily conserved DT-pairs are highly enriched in auto-regulatory transcription factors. Among the subset of

22 well-studied regulators, 17 have been annotated[39–41] to regulate their own biosynthesis (whereas less than half of the regulators not divergently transcribed in *E. coli* are auto-regulatory[39–41]). When analyzing the literature with respect to the remaining five proteins, we found three additional auto-regulators that have not been annotated in regulatory databases yet—thus, most, if not all transcriptional regulators encoded in conserved DT-pairs are auto-regulated (**Supplementary Notes** online).

Context-based function prediction using conserved DT-pairs

We use the described correlation for a novel gene context-based function prediction method, by requiring conservation of DT-pairs across at least three clades to gain confidence (this level of conservation is expected only if selected for). Specifically, we predict functional associations between bidirectionally oriented genes—including DT-pairs in species other than *E. coli*, for which predictions are transferred back to *E. coli* via orthology (a list with all conserved bidirectionally transcribed gene pairs is available online at http://www.bork.embl.de/Docu/Bidirectional_genes/). Altogether, 277 of such functional associations can be directly mapped to pairs of genes in the *E. coli* genome, a number comparable to the gene fusion method[1,2], which predicts 472 associations at the same level of conservation.

To quantify the accuracy of these predictions, pairs classified as RX or XX were independently benchmarked using annotated transcription regulatory interactions[39–41] and the KEGG (Kyoto encyclopedia of genes and genomes) database[42]. The lower limit for prediction accuracy (determined using a framework reported earlier[12]; see **Supplementary Notes** online) is 52% for RX and 34% for XX pairs, thus comparable to that of conserved codirectional gene neighbors (66%), and gene fusions (76%). The lower accuracy of our method is largely due to an enrichment of transcriptional regulators in poorly resolved orthologous groups[43] containing two or more similar regulators per species, which complicates transfer of function via orthology. For well-resolved[43] orthologous groups the accuracy for RX pairs improves to 75% (**Supplementary Notes** online); the predictor thus allows precise statements about the type and nature of the predicted association.

This can be extended to *de novo* function predictions for 'hypothetical' proteins encoded within DT-pairs, provided that a hypothetical protein can be classified as a transcriptional regulator by homology. For example, the three prokaryotic members of the poorly characterized orthologous group 'KOG2969' contained in STRING[12] are divergently oriented neighbors of *rpsU*, the gene encoding bacterial or mitochondrial ribosomal protein S2 (**Fig. 4**). Using PSI-BLAST[44], we detected homology with several members of the TetR family, a regulatory family previously described only in prokaryotes[23]. We predict that KOG2969 represents a novel bacterial-type regulatory family involved in regulating ribosome-associated genes in α-proteobacteria (*Caulobacter crescentus* and the genus *Rickettsia*) as well as in mitochondria, which arose from this bacterial group[45]. Recent large-scale experiments in yeast support a functional association between KOG2969 and mitochondrial RpsU: Ylr201c (Fmp53) is localized in the mitochondria[46], gene deletion leads to growth defects on a nonfermentable carbon source[47] (indicating an important role in this organelle) and its nuclear promoter is bound by the Abf1 transcription factor[48], which controls several ribosomal proteins[49]. We thus conclude that functional associations predicted by our method can be transferred to other species, even across the domains of life.

Considering all DT-pairs conserved across at least three clades in 101 genomes, we can predict 288 pairwise associations between orthologous groups[12,25] corresponding to 2,658 pairs of genes, out of which 63 associations between orthologous groups (671 between genes) are also covered by previously established genomic context approaches combined in STRING[12] (using a score cutoff of 0.400). The approach is thus complementary to the previous prediction methods based on gene fusion[1,2], the presence of genes in operons[3,4] and co-occurrence of genes across species[5] (see **Box 1** and **Fig. 5**).

Conclusions

Genomic context methods are widely used and continue to be improved; furthermore, new approaches are being reported (**Box 1** and references therein). Novel experimental data enable the detection of hitherto unknown signals; for example, global gene expression data indicate the coregulation of adjacent genes in prokaryotes regardless of relative orientation (**Box 2** and **Fig. 6**). In this article, we show that conserved bidirectionally transcribed gene neighbors (**Figs. 1**,**2**) are particularly strongly coregulated (**Fig. 3**) and can be exploited to make precise functional predictions. Namely, self-regulatory transcription factors can be associated to their target genes or operons, enabling hundreds of functional assignments.

As is the case for other genomic context methods, the number of predictions should increase as more genomes become sequenced. Furthermore, functional associations identified by the method can be transferred to eukaryotic orthologs of prokaryotic gene pairs—allowing (combined with homology searches) the characterization of hypothetical proteins (**Fig. 4**).

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS
This work was supported by the Bundesministerium für Forschung und Bildung, Germany. We would like to thank Aidan Budd, Toby Gibson, Florian Raible, David Ussery and members of the Bork group for helpful discussions—and in particular the former group members Martijn Huynen and Berend Snel for invaluable input in the early phase of this project.

COMPETING INTERESTS STATEMENT
The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturebiotechnology/

1. Enright, A.J., Iliopoulos, I., Kyrpides, N.C. & Ouzounis, C.A. Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**, 86–90 (1999).
2. Marcotte, E.M. *et al.* Detecting protein function and protein–protein interactions from genome sequences. *Science* **285**, 751–753 (1999).
3. Dandekar, T., Snel, B., Huynen, M. & Bork, P. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.* **23**, 324–328 (1998).
4. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D. & Maltsev, N. The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. USA* **96**, 2896–2901 (1999).
5. Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. & Yeates, T.O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA* **96**, 4285–4288 (1999).
6. Marcotte, E.M., Xenarios, I., van Der Bliek, A.M. & Eisenberg, D. Localizing proteins in the cell from their phylogenetic profiles. *Proc. Natl. Acad. Sci. USA* **97**, 12115–12120 (2000).
7. Kolesov, G., Mewes, H.W. & Frishman, D. SNAPping up functionally related genes based on context information: a colinearity-free approach. *J. Mol. Biol.* **311**, 639–656 (2001).
8. Mellor, J.C., Yanai, I., Clodfelter, K.H., Mintseris, J. & DeLisi, C. Predictome: a database of putative functional links between proteins. *Nucleic Acids Res.* **30**, 306–309 (2002).
9. Wu, J., Kasif, S. & DeLisi, C. Identification of functional links between genes using phylogenetic profiles. *Bioinformatics* **19**, 1524–1530 (2003).
10. Overbeek, R. *et al.* The ERGO genome analysis and discovery system. *Nucleic Acids Res.* **31**, 164–171 (2003).
11. Date, S.V. & Marcotte, E.M. Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nat. Biotechnol.* **21**, 1055–1062 (2003).
12. von Mering, C. *et al.* STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.* **31**, 258–261 (2003).
13. Salwinski, L. & Eisenberg, D. Computational methods of analysis of protein-protein interactions. *Curr. Opin. Struct. Biol.* **13**, 377–382 (2003).

14. Ouzounis, C.A., Coulson, R.M., Enright, A.J., Kunin, V. & Pereira-Leal, J.B. Classification schemes for protein structure and function. *Nat. Rev. Genet.* **4**, 508–519 (2003).
15. Valencia, A. & Pazos, F. Computational methods for the prediction of protein interactions. *Curr. Opin. Struct. Biol.* **12**, 368–373 (2002).
16. Aloy, P. & Russell, R.B. Interrogating protein interaction networks through structural biology. *Proc. Natl. Acad. Sci. USA* **99**, 5896–5901 (2002).
17. Jansen, R. *et al.* A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302**, 449–453 (2003).
18. Bader, G.D. *et al.* Functional genomics and proteomics: charting a multidimensional map of the yeast cell. *Trends Cell. Biol.* **13**, 344–356 (2003).
19. Alm, E. & Arkin, A.P. Biological networks. *Curr. Opin. Struct. Biol.* **13**, 193–202 (2003).
20. Li, S. *et al.* A map of the interactome network of the metazoan *C. elegans*. *Science* **303**, 540–543 (2004).
21. Bork, P. *et al.* Protein interaction networks from yeast to human. *Curr. Opin. Struct. Biol.* **14**, 292–299 (2004).
22. Altschul, S.F. & Koonin, E.V. Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases. *Trends Biochem. Sci.* **23**, 444–447 (1998).
23. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **30**, 276–280 (2002).
24. Letunic, I. *et al.* SMART 4.0: towards genomic data integration. *Nucleic Acids Res.* **32**, Database issue, D142–144 (2004).
25. Tatusov, R.L. *et al.* The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).
26. Morett, E. *et al.* Systematic discovery of analogous enzymes in thiamin biosynthesis. *Nat. Biotechnol.* **21**, 790–795 (2003).
27. Jacob, F. The operon after 25 years. *C.R. Acad. Sci. III* **320**, 199–206 (1997).
28. Salgado, H., Moreno-Hagelsieb, G., Smith, T.F. & Collado-Vides, J. Operons in *Escherichia coli*: genomic analyses and predictions. *Proc. Natl. Acad. Sci. USA* **97**, 6652–6657 (2000).
29. Rhee, K.Y. *et al.* Transcriptional coupling between the divergent promoters of a prototypic LysR-type regulatory system, the *ilvYC* operon of *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **96**, 14294–14299 (1999).
30. Adachi, N. & Lieber, M.R. Bidirectional gene organization: a common architectural feature of the human genome. *Cell* **109**, 807–809 (2002).
31. Beck, C.F. & Warren, R.A. Divergent promoters, a common form of gene organization. *Microbiol. Rev.* **52**, 318–326 (1988).
32. El-Robh, M.S. & Busby, S.J. The *Escherichia coli* cAMP receptor protein bound at a single target can activate transcription initiation at divergent promoters: a systematic study that exploits new promoter probe plasmids. *Biochem. J.* **368**, 835–843 (2002).
33. Stuart, J.M., Segal, E., Koller, D. & Kim, S.K. A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**, 249–255 (2003).
34. van Noort, V., Snel, B. & Huynen, M.A. Predicting gene function by conserved co-expression. *Trends Genet.* **19**, 238–242 (2003).
35. Huynen, M.A. & Snel, B. Gene and context: integrative approaches to genome analysis. *Adv. Protein Chem.* **54**, 345–379 (2000).
36. Bork, P. *et al.* Empirical and analytical approaches to gene order dynamics, map alignment and the evolution of gene families. in *Comparative Genomics*, vol. 1. (Sankoff, D. & Nadeau, J.H., eds.) 281–294 (Kluwer academic publishers, Dordrecht, 2000).
37. Gollub, J. *et al.* The Stanford Microarray Database: data access and quality assessment tools. *Nucleic Acids Res.* **31**, 94–96 (2003).
38. Zhou, X., Kao, M.C. & Wong, W.H. Transitive functional annotation by shortest-path analysis of gene expression data. *Proc. Natl. Acad. Sci. USA* **99**, 12783–12788 (2002).
39. Salgado, H. *et al.* RegulonDB (version 3.2): transcriptional regulation and operon organization in *Escherichia coli* K-12. *Nucleic Acids Res.* **29**, 72–74 (2001).
40. Munch, R. *et al.* PRODORIC: prokaryotic database of gene regulation. *Nucleic Acids Res.* **31**, 266–269 (2003).
41. Madan Babu, M. & Teichmann, S.A. Evolution of transcription factors and the gene regulatory network in *Escherichia coli*. *Nucleic Acids Res.* **31**, 1234–1244 (2003).
42. Kanehisa, M., Goto, S., Kawashima, S. & Nakaya, A. The KEGG databases at GenomeNet. *Nucleic Acids Res.* **30**, 42–46 (2002).
43. Von Mering, C. *et al.* Genome evolution reveals biochemical networks and functional modules. *Proc. Natl. Acad. Sci. USA* **100**, 15428–15433 (2003).
44. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
45. Gabaldon, T. & Huynen, M.A. Reconstruction of the proto-mitochondrial metabolism. *Science* **301**, 609 (2003).
46. Huh, W.K. *et al.* Global analysis of protein localization in budding yeast. *Nature* **425**, 686–691 (2003).
47. Steinmetz, L.M. *et al.* Systematic screen for human disease genes in yeast. *Nat. Genet.* **31**, 400–404 (2002).
48. Lee, T.I. *et al.* Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298**, 799–804 (2002).
49. Warner, J.R. The economics of ribosome biosynthesis in yeast. *Trends Biochem. Sci.* **24**, 437–440 (1999).
50. Snel, B., Bork, P. & Huynen, M.A. Genome phylogeny based on gene content. *Nat. Genet.* **21**, 108–110 (1999).
51. Korbel, J.O., Snel, B., Huynen, M.A. & Bork, P. SHOT: a web server for the construction of genome phylogenies. *Trends Genet.* **18**, 158–162 (2002).
52. Hedges, S.B. The origin and evolution of model organisms. *Nat. Rev. Genet.* **3**, 838–849 (2002).
53. Feng, D.F., Cho, G. & Doolittle, R.F. Determining divergence times with a protein clock: update and reevaluation. *Proc. Natl. Acad. Sci. USA* **94**, 13028–13033 (1997).
54. Doolittle, R.F., Feng, D.F., Tsang, S., Cho, G. & Little, E. Determining divergence times of the major kingdoms of living organisms with a protein clock. *Science* **271**, 470–477 (1996).
55. Workman, C. *et al.* A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biol.* **3**, research0048, 30 August 2002, doi:10.1186/gb-2002-3-9-research0048.
56. Huelsenbeck, J.P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755 (2001).
57. Chenna, R. *et al.* Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* **31**, 3497–3500 (2003).
58. Tatusov, R.L., Koonin, E.V. & Lipman, D.J. A genomic perspective on protein families. *Science* **278**, 631–637 (1997).
59. Huynen, M.A. & Bork, P. Measuring genome evolution. *Proc. Natl. Acad. Sci. USA* **95**, 5849–5856 (1998).
60. Marcotte, E.M. Computational genetics: finding protein function by nonhomology methods. *Curr. Opin. Struct. Biol.* **10**, 359–365 (2000).
61. Galperin, M.Y. & Koonin, E.V. Who's your neighbor? New computational approaches for functional genomics. *Nat. Biotechnol.* **18**, 609–613 (2000).
62. Osterman, A. & Overbeek, R. Missing genes in metabolic pathways: a comparative genomics approach. *Curr. Opin. Chem. Biol.* **7**, 238–251 (2003).
63. Huynen, M.A., Snel, B., von Mering, C. & Bork, P. Function prediction and protein networks. *Curr. Opin. Cell. Biol.* **15**, 191–198 (2003).
64. Pazos, F. & Valencia, A. Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng.* **14**, 609–614 (2001).
65. Thomas, G., Coutts, G. & Merrick, M. The glnKamtB operon. A conserved gene pair in prokaryotes. *Trends Genet.* **16**, 11–14 (2000).
66. Coutts, G., Thomas, G., Blakey, D. & Merrick, M. Membrane sequestration of the signal transduction protein GlnK by the ammonium transporter AmtB. *EMBO J.* **21**, 536–545 (2002).
67. Weller, G.R. *et al.* Identification of a DNA nonhomologous end-joining complex in bacteria. *Science* **297**, 1686–1689 (2002).
68. Daugherty, M., Vonstein, V., Overbeek, R. & Osterman, A. Archaeal shikimate kinase, a new member of the GHMP-kinase family. *J. Bacteriol.* **183**, 292–300 (2001).
69. Huynen, M.A., Snel, B., Bork, P. & Gibson, T.J. The phylogenetic distribution of frataxin indicates a role in iron-sulfur cluster protein assembly. *Hum. Mol. Genet.* **10**, 2463–2468 (2001).
70. Muhlenhoff, U., Richhardt, N., Ristow, M., Kispal, G. & Lill, R. The yeast frataxin homolog Yfh1p plays a specific role in the maturation of cellular Fe/S proteins. *Hum. Mol. Genet.* **11**, 2025–2036 (2002).
71. Myllykallio, H. *et al.* An alternative flavin-dependent mechanism for thymidylate synthesis. *Science* **297**, 105–107 (2002).
72. Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* **3**, 318–356 (1961).
73. Sabatti, C., Rohlin, L., Oh, M.K. & Liao, J.C. Co-expression pattern from DNA microarray experiments as a tool for operon prediction. *Nucleic Acids Res.* **30**, 2886–2893 (2002).
74. Zheng, Y., Szustakowski, J.D., Fortnow, L., Roberts, R.J. & Kasif, S. Computational identification of operons in microbial genomes. *Genome Res.* **12**, 1221–1230 (2002).
75. Chandler, M.G. & Pritchard, R.H. The effect of gene concentration and relative gene dosage on gene output in *Escherichia coli*. *Mol. Gen. Genet.* **138**, 127–141 (1975).
76. Ehira, S., Ohmori, M. & Sato, N. Genome-wide expression analysis of the responses to nitrogen deprivation in the heterocyst-forming cyanobacterium *Anabaena* sp. strain PCC 7120. *DNA Res.* **10**, 97–113 (2003).
77. Hatfield, G.W. & Benham, C.J. DNA topology-mediated control of global gene expression in *Escherichia coli*. *Annu. Rev. Genet.* **36**, 175–203 (2002).
78. Dorman, C.J. & Deighan, P. Regulation of gene expression by histone-like proteins in bacteria. *Curr. Opin. Genet. Dev.* **13**, 179–184 (2003).