Computational approaches for the analysis of gene neighbourhoods in prokaryotic genomes

Igor B. Rogozin, Kira S. Makarova, Yuri I. Wolf and Eugene V. Koonin

Date received (in revised form): 15th March 2004

Igor B. Rogozin
is a staff scientist at the National Center for Biotechnology Information NLM/NIH (Bethesda, MD, USA) and a research scientist at the Institute of Cytology and Genetics RAS (Novosibirsk, Russia).

Kira S. Makarova and **Yuri I. Wolf**
are staff scientists at the National Center for Biotechnology Information NLM/NIH (Bethesda, MD, USA).

Eugene V. Koonin
is a group leader at the National Center for Biotechnology Information NLM/NIH (Bethesda, MD, USA).

Keywords: *gene order, operon, phylogenetic analysis, functional signal, local alignment, dot-plot, conserved gene pair*

Abstract

Gene order in prokaryotes is conserved to a much lesser extent than protein sequences. Only some operons, primarily those that encode physically interacting proteins, are conserved in all or most of the bacterial and archaeal genomes. Nevertheless, even the limited conservation of operon organisation that is observed provides valuable evolutionary and functional clues through multiple genome comparisons. With the rapid growth in the number and diversity of sequenced prokaryotic genomes, functional inferences for uncharacterised genes located in the same conserved gene neighborhood with well-studied genes are becoming increasingly important. In this review, we discuss various computational approaches for identification of conserved gene strings and construction of local alignments of gene orders in prokaryotic genomes.

THE UNITS OF GENOME ORGANISATION IN PROKARYOTES: OPERONS, GENE PAIRS AND DIRECTONS

Study of gene location in the genome is one of the classic areas of genetics. Non-random associations between genes have been observed ever since the first genetic maps were constructed, and explicit analysis of genome rearrangements was pioneered by Dobzhansky and Sturtevant in the 1930s.[1] However, sequencing of numerous prokaryotic and eukaryotic genomes during the 1990s put analysis of gene order on a new footing by allowing direct and comprehensive comparative analysis of gene locations in genomic sequences. The biological significance and evolutionary dynamics of gene co-localisation are substantially different in prokaryotes and eukaryotes. Operons, groups of adjacent, co-expressed genes that often encode functionally linked proteins, represent the principal form of gene co-regulation in prokaryotes (Figure 1).[2–4] Some of the operons, particularly those that encode subunits of multiprotein complexes, such as ribosomal proteins, are shared by phylogenetically distant bacterial genomes or even between archaea and bacteria (Figure 2).[5–7] This is due, in part, to conservation of these operons over long stretches of evolutionary time, perhaps even since the last universal common ancestor of all modern life forms, and, in part, to horizontal spread of operons among prokaryotes.[8] In eukaryotes, operons have been detected in nematodes and some protists,[9] but most eukaryotic genes form autonomous transcription units and are expressed largely independently from each other.[10]

As discussed below in more detail, information about co-localised prokaryotic genes (gene neighbourhoods) can be used for functional inferences. Prediction of functional coupling between genes is based on conservation of gene clusters between genomes. If the function of one gene in a conserved gene cluster is known, the function of a

B. Rogozin,
NCBI/NLM/NIH,
8600 Rockville Pike,
Bldg. 38A,
Bethesda, MD 20894, USA

Tel: +1 301 594 4271
Fax: +1 301 435 7794
E-mail: rogozin@ncbi.nlm.nih.gov

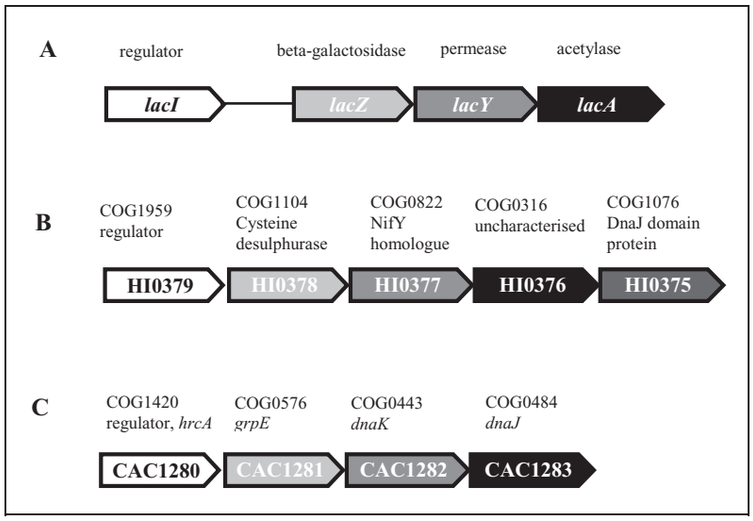

Figure 1: Examples of (predicted) operons with adjacent genes coding for the respective regulators. (A) The *Escherichia coli lacZYA* operon; (B) a predicted operon of *Haemophilus influenzae* consisting of genes potentially involved in assembly of redox protein complexes; (C) predicted heat shock protein operon of *Clostridium acetobutilicum*. In each case, the (predicted) DNA-binding regulator is encoded by the upstream gene. Genes are shown not to scale; the direction of transcription is indicated by arrows

Functional inferences

neighbouring gene (presumably from the same operon) can be inferred through the 'guilt by association' principle.[11] For example, co-localisation of a predicted transcriptional regulator (COG1959) with genes potentially involved in assembly of redox protein complexes (Figure 1B) in several prokaryotic genomes suggests that this regulator specifically modulates the expression of genes for the subunits of these complexes. The utility of gene neighbourhood analysis for functional prediction can be further augmented when this analysis is combined with analyses of gene fusion events, expression arrays data, protein–protein interactions, metabolic pathways, phylogenetic profiles and other aspects of genomic context.[11–20]

Directions

A pair of adjacent genes is a natural unit of gene co-localisation. There are three types of gene pairs with respect to the directions of transcription: (i) unidirectional, (ii) convergent and (iii) divergent (Figure 3). The three classes of spacers defined by these distinct gene arrangements differ in terms of the types of regulatory sites they contain. Spacers between unidirectional genes may include both a terminator for the upstream gene, and a promoter and additional signals, such as transcription factor binding sites and enhancers/silencers, for the downstream gene; spacers between convergent genes contain exclusively terminators; and spacers between divergent genes have only promoters and other upstream transcriptional signals. In prokaryotes, regions separating adjacent unidirectional genes represent a mixture of inter- and intra-operonic spacers, whereas convergent and divergent gene pairs contain only inter-operonic spacers. A clear peak at short distances between genes in the same operon contrasts a flat length distribution of inter-operonic distances, and this property was used for predicting operons in *Escherichia coli*.[21] For this purpose, sets of genes transcribed in the same direction, with no intervening gene transcribed in the opposite direction, and separated by relatively short non-coding spacers, were clustered into 'directons'. This straightforward approach yielded 812 directons with more than one gene. Despite its remarkable simplicity, the directon approach correctly identified ~75 per cent of the known operons in the *E. coli* genome and therefore seems to be a reasonably reliable method for prediction of operons.[21,22]

CENTRAL DEFINITIONS OF GENE NEIGHBOURHOOD ANALYSIS

Orthologous genes

Orthologous and pathologous genes

To compare gene orders in different genomes, one needs, first, to establish orthologous (or, in simpler analysis schemes, homologous) relationships between genes. Orthologues are defined as homologous genes that derive by vertical descent from a single ancestral gene in the last common ancestor of the compared species. Paralogues, in contrast, are homologous genes that, at some stage of evolution of the respective gene family, have evolved by duplication of an ancestral gene.[23–25] Orthologous genes

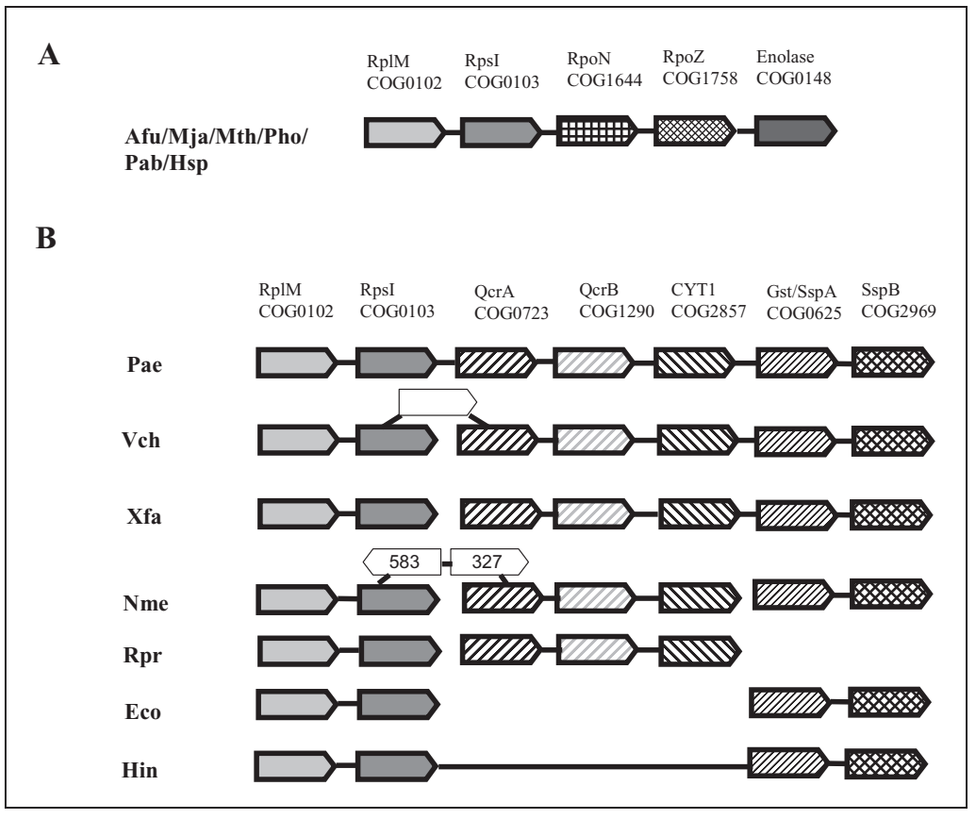

Figure 2: Fragments of a ribosomal protein gene neighbourhood containing apparent hitchhiker genes. Shaded or hatched arrows indicate COGs that belong to the ribosomal protein gene neighbourhood; empty arrows indicate inserted genes. Orthologous genes are shown by the same pattern. (A) The gene for the glycolytic enzyme enolase is part of the ribosomal protein gene cluster in Euryarchaeota. COG0102, large subunit ribosomal protein L13; COG0103, small subunit ribosomal protein S9; COG1644, DNA-directed RNA polymerase, subunit N; COG1758, DNA-directed RNA polymerase, subunit K; COG0148, enolase. (B) Proteobacterial ribosomal protein cluster includes genes for stringent starvation response proteins, which appear to be functionally linked to translation, and genes for electron transfer chain components, probable hitchhikers. COGs absent in (A): COG0723, Rieske Fe-S cluster protein; COG1290, cytochrome b subunit of the bc complex; COG2857, cytochrome c1; COG0625, stringent starvation protein A (glutathione *S*-transferase); COG2969, stringent starvation protein B; COG0583, transcriptional regulator; COG0327, uncharacterised conserved protein. Genes are shown not to scale; the direction of transcription is indicated by arrows.
Abbreviations: Afu *Archaeoglobus fulgidus*; Mja *Methanococcus jannaschii*; Mth *Methanobacterium thermoautotrophicum*; Pho *Pyrococcus horikoshii*; Pab *Pyrococcus abyssi*; Hsp *Halobacterium* sp.; Pae *Pseudomonas aeruginosa*; Vch *Vibrio cholerae*; Xfa *Xylella fastidiosa*; Nme *Neisseria meningitidis*; Rpr *Rickettsia prowazekii*; Eco *Escherichia coli*; Hin *Haemophilus influenzae*

Clusters of Orthologous Groups (COGs)

can be identified by various means. One approach is to use existing databases of orthologues, eg Clusters of Orthologous Groups of proteins (COGs) (Table 1).[26–28] The COGs were constructed from the results of all-against-all BLAST comparison of proteins encoded in complete genomes by detecting consistent sets of genome-specific best hits (BeTs). The COG construction procedure did not rely on any preconceived phylogenetic tree of the included species except that certain obviously related genomes (for example, two species of mycoplasmas or pyrococci) were grouped prior to the analysis, to eliminate strong

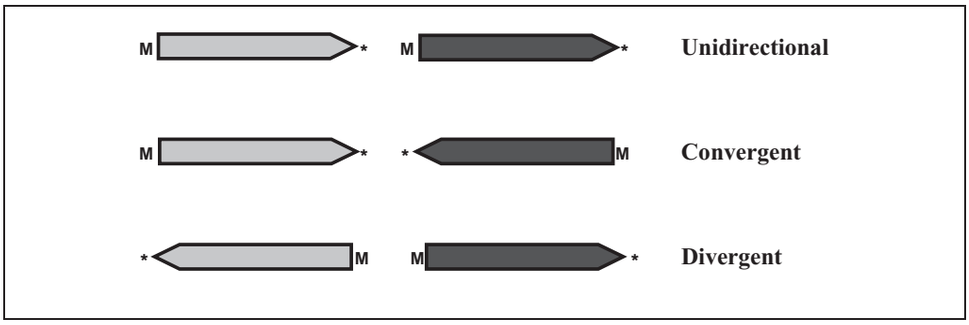

Figure 3: Three types of gene pairs. M, N-terminal methionine; *, stop-codon. The direction of transcription is indicated by arrows

Reciprocal BLAST best hits

dependence between BeTs.[26] Another approach to identification of probable orthologues involves using pairwise genome comparisons and detecting pairs of orthologous genes as reciprocal BeTs. Both these methods are predicated on the assumption that orthologous genes are most similar among all compared pairs of genes.[6,29–31] The reciprocal BeT approach is less reliable but has the advantage of being applicable to any pair of newly sequenced genomes regardless of whether or not they have already been included in the COGs or other similar databases.

Functionally related and evolutionarily conserved gene clusters

Conserved gene clusters

When considering gene co-localisation in genomes, one should distinguish between two connected but distinct conceptual frameworks: clusters of functionally related genes and evolutionarily conserved gene clusters. The former concept deals with functions of genes and therefore involves a degree of arbitrariness, especially in cases when gene functions have not been characterised in detail. The latter notion is based on identification of conserved gene strings in distantly related genomes and is more amenable to formal treatment. An additional important distinction exists between gene strings shared by two or more genomes of relatively close species simply because there was not enough time since their divergence from a common ancestor for recombination to obliterate the ancestral gene order, and functionally important gene strings that are maintained by purifying selection. Functional inferences based on gene order conservation are legitimate only when the analysed genomes are completely 'saturated' by recombination events. There is little doubt that such saturation had been already reached in the case of taxonomically distant species, such as archaea and bacteria. However, for more closely related genomes, such as those of different bacteria from the same lineage, the exact boundary at which purely

Table 1: Web-servers for gene order analysis and other relevant sites

Name	Address
STRING	www.bork.embl-heidelberg.de/STRING/
SNAPPER	pedant.gsf.de/snapper/
RegulonDB	www.cifn.unam.mx/Computational_Genomics/regulondb/
LAMARCK gene strings	ftp.ncbi.nlm.nih.gov/pub/koonin/genome_align/
COGs	www.ncbi.nlm.nih.gov/COG/
KEGG	www.genome.ad.jp/kegg/
EcoCyc	ecocyc.org/
ERGO	http://ergo.integratedgenomics.com/ERGO/
Missing genes, genome context analysis	www.integratedgenomics.com/online_material/osterman/index.html
SHOT	http://www.bork.embl-heidelberg.de/~korbel/SHOT/

'historical' conservation of gene order becomes negligible is hard, if not impossible, to determine.

In this review, we discuss approaches to gene order analysis that are based on both the functional and the evolutionary aspects of gene clustering in genomes. Several explanations for clustering of functionally related genes have been proposed:[8,9]

Co-regulation of expression

- The co-regulation model postulates that clustering of functionally related genes is maintained by selection because co-regulation of their expression from a single promoter is beneficial for the organism. This model is implicit in the original description of operons, which was developed primarily from studies on the genes encoding lactose utilisation in *Escherichia coli*, the *lacZYA* operon (Figure 1A).[2–4] These genes are co-regulated, being induced by lactose, and this principle works for many other operons, eg those that encode proteins involved in utilisation of other nutrients.

- The Natal model proposes that clusters of functionally related genes originate by tandem duplications.

- The Fisher model postulates that co-localisation of co-adapted genes reduces the frequency of deleterious recombination events disrupting these complexes.

- The molarity model proposes that gene clustering results in beneficial high local concentrations of interacting proteins.

Selfish operon hypothesis

- The selfish operon model suggests that operons escape elimination by invasion of new genomes.

According to this last concept, gene clusters behave similarly to selfish genetic elements, such as transposons, with the clustering being initially beneficial to the genes themselves, not to the host organisms.[8,32] Genes that form a cluster obviously have a greater chance than dispersed genes to propagate via joint horizontal transfer; according to the selfish operon hypothesis, this simple fact, rather than any functional adaptations, is the main cause of the long-term survival of operons. This concept most readily applies to gene clusters, which have no essential functions but are advantageous under specific conditions, eg acquisition of the *lacZYA* operon is beneficial for bacteria in a lactose-containing medium. The selfish operon hypothesis is not incompatible with the co-regulation concept. Indeed, it seems that the two can be easily reconciled by postulating that horizontal transfer of an entire operon is favoured by selection over transfer of individual genes because, in the former case, gene co-expression and co-regulation are preserved.[33] Dissemination of operons via horizontal transfer appears particularly plausible for operons (including *lacZYA),* in which the regulator is encoded next to the regulated genes (Figure 1). However, such operons are relatively uncommon, and for those operons that are not adjacent to the regulator gene, preservation after horizontal transfer becomes problematic owing to probable absence of a compatible regulator in the recipient organism. In general, it appears that a combination of co-regulation with the selfish operon mechanism provides the most plausible explanation for the wide spread of operons in prokaryotes, whereas the factors emphasised by other hypotheses are of minor importance at best.

CONCEPTS AND METHODS OF GENE NEIGHBOURHOOD ANALYSIS

General principles and problems

Gene order at a level above operons is poorly conserved, and genome comparison diagonal plots, in which points indicate orthologues, appear completely disordered even for species

that belong to the same prokaryotic lineage, eg *E. coli and Haemophilus influenzae*, two members of the gamma-subdivision of Proteobacteria.[5,29] Operons, which typically consist of three to four genes, tend to be substantially more conserved in evolution than non-operonic gene strings. Hence the two modes of gene order evolution in prokaryotes: the constrained and, consequently, relatively slow intra-operon gene rearrangement as opposed to the rapid shuffling of operons.

Operons tend to be more conserved than non-specific gene strings

Comparative analysis of gene orders is complicated by numerous errors in genome annotations, the most common ones being incorrect assignment of translations starts, falsely predicted genes and missed genes, and frameshifts.[34–37] Furthermore, many parasitic bacteria, eg *Mycobacterium leprae* and *Rickettsia prowazekii*, have numerous pseudo-genes in their genomes,[38–40] which may be hard to recognise, resulting in ambiguities in errors in identification of orthologues. Given these and other problems, upon which we will touch in the more technical discussion below, it is not surprising that, despite sustained effort of many research groups over several years, there is so far no single satisfactory strategy for comparative analysis of gene orders in prokaryotes. In this review, we critically discuss the principal computational approaches employed gene order analysis and illustrate their achievements with biologically important results of comparative genomics.

Errors of genome annotation, missed genes, frameshifts

Comparison of gene orders in genomes bears obvious similarities to the more familiar comparison of nucleotide and amino acid sequences. The irony here is that, historically, gene order analysis was incepted before the very idea of a molecular sequence came to the fore;[1] however, in the genomic era, computational methods for sequence comparison obviously took the driver's seat. In each of these situations, the basic procedure involves comparison and alignment of strings of symbols drawn from a fixed alphabet by using a chosen scoring system (1 for a match and 0 for a mismatch in the simplest case). Apart from this central common theme, however, there are substantial differences between the comparison of molecular sequences and gene orders. In sequence analysis, the alphabet is small and universal: 4 bases in nucleic acids and 20 amino acids in proteins. By contrast, in gene order analysis, the alphabet is typically large and unique for each pair of compared genomes because usually it consists of all orthologous genes. Detection of orthologues itself depends on sequence comparison, typically, of protein sequences. Thus, gene order analysis is naturally viewed as a meta-procedure with respect to sequence analysis and, accordingly, in addition to its own problems and caveats, inherits those of sequence comparison. The second major difference between sequence analysis and gene order analysis is that, in the former, sequences are normally treated as collinear; permutations are rare and special procedures to handle them are not deemed critically important. In contrast, at the genome level, gene shuffling resulting in numerous permutations is extremely common and cannot possibly be disregarded in any comparison procedure. Hence, comparison of gene orders presents challenges above and beyond those that are already familiar to computational biologists from the experience of sequence analysis.

Global and local alignments

Theoretically, analysis of gene order conservation is similar to alignment of other biological sequences (DNA or protein) in that either global or local alignments can be constructed.[41,42] However, the global alignment approach, which is not practicable even for distantly related protein sequences, is not applicable to prokaryotic genome alignment at all (except, possibly, pairs of very closely related isolates of the same microbial strain) owing to a large number of deletions, insertions, translocations and inversions that occur during evolution.

The evolutionary fluidity of

prokaryotic chromosomes makes comparative analysis of gene order in distantly related genomes a non-trivial task. Several methods have been proposed and tested for comparing gene orders in pairs of genomes and in multiple genomes, and detecting local gene order conservation; these methods differ in the amount of gene insertion/deletion and local rearrangement that they allow.

Genomic dot-plots and alignments

Genomic global pairwise alignment of gene orders

A simple way of comparing two sequences[43] or two gene orders[29] is to construct a dot-plot. Examples of genomic dot-plots are shown in Figure 4. In these comparisons, each dot corresponds to a pair of orthologous genes and the projections of a dot on the two axes are the respective locations of these orthologues in the compared genomes. If the two genomes are fully collinear, there will be a perfect diagonal line of dots. In practice, however, even closely related species have undergone various recombination events (deletions, insertions, duplications, translocations and inversions) since the time of their divergence from the common ancestor, which causes deviations from the diagonal pattern (Figure 4A,B,C). As an alternative to depiction of orthologous genes only, it is possible to include in dot-plot analysis all gene pairs that show sequence similarity above a selected threshold. Obviously, this yields, in general, a richer, more complicated picture because many paralogous genes and even gene strings are detected (Figure 4D). The genomic dot-plot approach is particularly useful for comparative analysis of gene orders in not too distantly related species.[44–46] Dot-plots of the conserved protein sequences between each pair of such species produce a distinct X-shaped pattern (Fig. 4), which was dubbed an X-alignment.[45] The key feature of these alignments is that they show symmetry around the replication origin and terminus; that is, the distance between particular conserved gene and the replication origin (or terminus) is conserved between closely related species. This suggests that large chromosomal inversions reversed the genomic sequence symmetrically around the origin of replication; such symmetrical inversions appear to be a common feature of bacterial genome evolution.[44–46]

Pairwise and template-based multiple alignment of gene orders

Itoh and coworkers compared gene orders in known operons from *E. coli* and *Bacillus subtilis* with corresponding gene strings in 11 complete genome sequences.[47] The degree of variation depended on the genomes examined, notably increasing with the increase of the evolutionary distance between the respective bacteria. It was suggested that shuffling of gene order was virtually neutral in long-term evolution.[47]

KEGG database

Kanehisa and coworkers combined information on biochemical pathways extracted from the KEGG database[48] (Table 1) with pairwise and multiple alignments of gene orders.[17,49] Their approach to the construction of such alignments is based on graph analysis: the genome was transformed into a graph with genes as nodes, and the pathway was represented as a separate graph with gene products as nodes. A simple method was developed to identify local similarities between two graphs (correlated clusters), allowing for gaps and mismatches of nodes and edges.[49] This method was applied to a comparison of completely sequenced genomes and the KEGG metabolic pathways. A tendency for formation of correlated clusters called FRECs (Functionally Related Enzyme Clusters) was revealed. However, this tendency varied considerably depending on the organism. The fraction of enzymes included in FRECs was close to 50 per cent for *B. subtilis* and *E. coli*, but was below 10 per cent for *Synechocystis*. The information from FRECs was used to refine orthologue group tables in KEGG.[17,49] A similar approach for

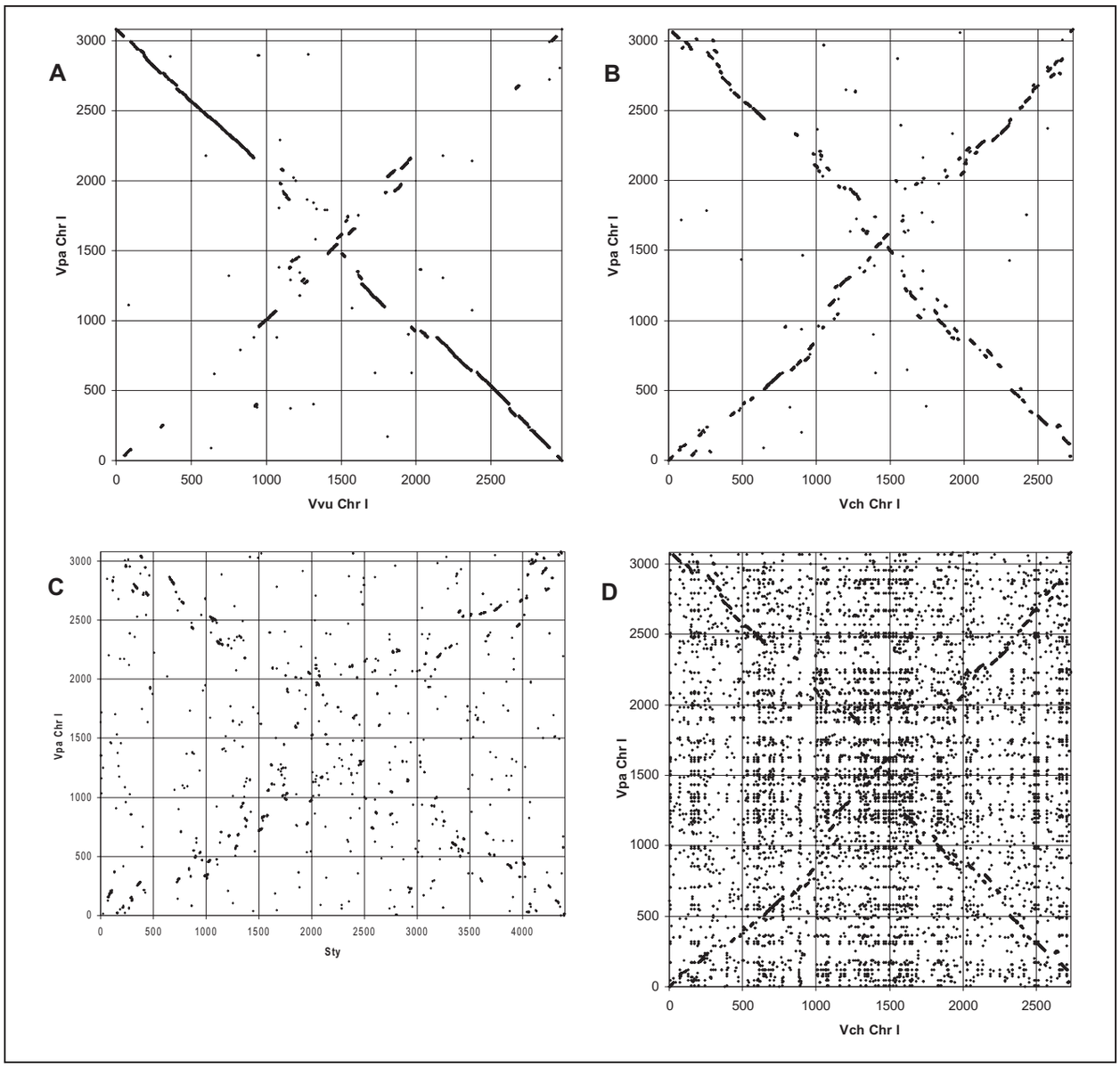

Figure 4: Dot-plot comparisons of prokaryotic genomes. (A) *Vibrio parahaemolyticus v. Vibrio vulnificus*, orthologous genes; (B) *Vibrio parahaemolyticus v. Vibrio cholerae*, orthologous genes; (C) *Vibrio parahaemolyticus v. Salmonella typhimurium*, orthologous genes; (D) *Vibrio parahaemolyticus v. Vibrio cholerae*, orthologous and paralogous genes

5.25 per cent of the genes belong to conserved gene strings

prediction of operons was developed by Zheng and coworkers.[18]

A systematic analysis of gene order conservation among prokaryotes was performed using the LAMARCK program, which constructs gapped local alignments of gene orders (Table 1); the statistical significance of the produced local alignments was assessed using Monte Carlo simulations.[30] This study showed that only 5–25 per cent of the genes in bacterial and archaeal genomes belong to gene strings (probable operons) shared by at least two genomes, once closely related species are excluded. Sets of local alignments were generated for all pairs of completely sequenced bacterial and archaeal genomes, and for each genome a so-called template-anchored multiple alignment was constructed. In this type of alignment,

Ribosomal superoperon contains >50 genes

each gene in the template genome is overlaid with the orthologous genes from the local alignments (conserved gene strings) identified in each of the other analysed genomes. This seems to be the best attainable surrogate for a multiple alignment of gene orders given the pervasive permutation problem (Figure 5). The majority of the conserved gene strings detected during this analysis were previously identified operons, with the ribosomal superoperon being the top-scoring string in most genome comparisons. However, in some of the bacterial–archaeal pairs, the superoperon is rearranged to the extent that other operons, primarily those subject to horizontal transfer, show the greatest level of conservation, such as the archaeal-type H+-ATPase operon or ABC-type transport cassettes. The potential of using template-anchored multiple-genome alignments for predicting functions of uncharacterised genes was quantitatively assessed. Functions were predicted or significantly clarified for approximately 90 COGs (~4 per cent of the total of 2,414 analysed COGs). The most significant

E.coli		bsu	mtu	hin	nme	rpr	hpy	cje	syn	dra	aae	tma	bbu	tpa	cpn	ctr	mpn	mge	uur	aer	afu	pyr	mja	mth
3642	bglB	•																						
3643	bglF	•																						
3644	bglG																							
3645	phoU									•		•						•	•		•			•
3646	pstB	•							•	•		•	•					•	•	•	•		•	•
3647	pstA	•	•	•					•	•	•	•	•					•	•	•	•		•	•
3648	pstC	•	•	•					•	•	•	•	•							•	•		•	•
3649	pstS	•	•	•					•	•	•	•								•			•	•
3652	atpC	•	•	•	•	•	•	•	•			•												
3653	atpD	•	•	•	•	•	•	•	•		•	•					•	•	•					
3654	atpG	•	•	•	•	•	•	•	•		•	•					•	•	•					
3655	atpA	•	•	•	•	•	•	•	•			•					•	•						
3656	atpH	•	•	•	•	•		•	•			•					•	•						
3657	atpF	•	•	•	•			•	•			•					•	•						
3658	atpE	•	•	•	•				•			•					♦	♦						
3659	atpB	•	•	•	•				•			•					•	•						
3660	atpI			♦																				
3668	kup			♦																				
3669	rbsD	•		•																				
3670	rbsA	•		•								•												
3671	rbsC	•		•								•												
3672	rbsB	•		•																				
3673	rbsK			•																				
3674	rbsR			•																				

Figure 5: Segments of a template-anchored, gene-by-gene genome alignment (template *Escherichia coli*). The first column shows the position of the respective gene in the template genome (*E. coli* genes were numbered from 1 to 4,279) and the second column shows the gene name. Black circles in the rest of the columns show the presence of the respective gene string in the corresponding pairwise genome comparison. Black diamonds indicate positions with gaps or mismatches in the gene strings.
Abbreviations: bsu *Bacillus subtilis*, mtu *Mycobacterium tuberculosis*, hin *Haemophilus influenzae*, nme *Neisseria meningitidis*, rpr *Rickettsia prowazekii*, hpy *Helicobacter pylori*, cje *Campylobacter jejunii*, syn *Synechocystis* PCC6803, dra *Deinococcus radiodurans*, aae *Aquifex aeolicus*, tma *Thermotoga maritima*, bbu *Borrelia burgdorferi*, tpa *Treponema pallidum*, cpn *Chlamydophila pneumoniae*, ctr *Chlamydia trachomatis*, mpn *Mycoplasma pneumoniae*, mge *Mycoplasma genitalium*, uur *Ureaplasma urealyticum*, aer *Aeropyrum pernix*, afu *Archaeoglobus fulgidus*, pyr *Pyrococcus abyssi*, mja *Methanococcus jannaschii*, mth *Methanobacterium thermoautotrophicum*

predictions were obtained for the poorly characterised archaeal genomes; these included a previously uncharacterised restriction-modification system, a nuclease-helicase combination implicated in DNA repair, and the probable archaeal counterpart of the eukaryotic exosome.[30,50] The latter prediction has been recently validated by experimental detection of an exosome-like complex in the archaeon *Sulfolobus solfataricus*.[51]

Gene context

STRING server

A single gene can be used as a query to study recurrent presence of other genes in a surrounding neighbourhood in multiple genomes. Snel and coworkers[52] developed STRING (Search Tool for Recurring Instances of Neighboring Genes) (Table 1), a tool to retrieve and display the genes with which a query gene repeatedly co-occurs. This tool employs the COG database and additional, unsupervised sets of putative orthologues as the source of information on gene conservation, and performs iterative search for recurring genomic neighbourhoods. The resulting genomic context of the query gene is integrated with additional information on its phyletic profile and visualised in several formats.[52]

Conserved gene pairs

Conserved gene arrays are a good approximation of actual operons

Systematic comparisons of bacterial and archaeal genomes revealed numerous conserved pairs of adjacent genes.[7,30,33,53–55] In some studies, a pair of genes was considered to be evolutionarily conserved if the respective genes were transcribed in the same direction and separated by zero, one or two genes. This relaxed definition of a conserved gene pair was adopted because numerous rearrangements, deletions and insertions have been found in operons whose characteristic size is three to five genes.[6,30] Several lines of evidence suggested that conserved pairs of unidirectional genes belong to conserved operons.[7,30,33,53–56] Firstly, very few, if any, conserved pairs of convergent or divergent genes were detected in phylogenetically distant genomes; this is most compatible with the notion that conservation of many unidirectional pairs has to do with co-expression and co-regulation.[5,7,21,30,33,53–56] Secondly, short distances that typically separate genes in conserved unidirectional pairs are in good agreement with this hypothesis because short spacers are usually observed within operons.[21,53,57] The distribution of distances between conserved unidirectional gene pairs in *E. coli*[57] was compared with the distribution of distances between genes in documented *E. coli* operons from the RegulonDB database[58] (Table 1). There was no significant difference between the two distributions; furthermore, none of the conserved gene pairs belonged to different documented *E. coli* operons and, for 81 per cent of the conserved gene pairs, both genes belonged to the same documented operon. These observations suggest that the set of conserved gene pairs is a good approximation of the set of genes from actual operons.[57]

Analysis of gene neighbourhoods in prokaryotic genomes based on conserved gene pairs

Über-operon concept

Bork and coworkers proposed the concept of 'über-operon', a set of genes whose functional and regulatory contexts tend to be conserved despite numerous rearrangements.[59] The conglomerate of operons encoding ribosomal proteins, the largest group of genes whose order is partially conserved in all prokaryotic genomes, is the paragon of an über-operon. It has to be emphasised that an über-operon does not necessarily portray the arrangement of the given set of genes in any extant or ancestral genome; instead, the composition and order of genes included in an über-operon seem to reflect multiple, alternative pathways of evolution.[59]

By combining the pairwise interactions between proteins, as predicted by the conserved co-occurrence of the respective genes in operons, Snel and coworkers

built networks of relationships between proteins.[60] The complete network contained 3,033 orthologous protein sets from 38 genomes. The network consisted of one giant component, containing 1,611 genes, and of 516 small clusters that, on average, contained only 2.7 genes. These small clusters had a homogeneous functional composition and thus apparently represented functional modules. Analysis of the giant component revealed that it was a scale-free, small-world network with a high degree of local clustering. It consisted of locally highly connected subclusters joined by linker proteins. The linker proteins tended to have multiple functions or were involved in multiple processes and had an above average probability of being essential. By splitting up the giant component at these linker proteins, 265 subclusters that tended to have a homogeneous functional composition were identified.[60]

Gene neighbourhoods

Rogozin and coworkers further developed the über-operon concept by delineating extended gene neighbourhoods by combining the results gene order comparisons in multiple prokaryotic genomes.[33] A flow chart of the procedure for construction of gene neighbourhoods from conserved gene pairs is shown in Figure 6. The idea behind this approach is that different genomes contain different, overlapping parts of evolutionarily and functionally connected gene neighbourhoods and, by generating a 'tiling path' through these overlaps, the entire neighbourhood can be reconstructed (Figure 7). The comparative-genomic approach used in this work was deliberately inclusive and aimed at detection of large, complex gene neighbourhoods. Accordingly, many of the resulting objects are complicated conglomerates of numerous, overlapping gene arrays. Most of these arrays, let alone the larger neighbourhoods, are not represented, in their entirety, in any particular genome. The very fact that the detected neighbourhoods are branched structures consisting of overlapping gene arrays indicates that they are neither reconstructions of an ancestral gene order nor functional domains in any particular genome, although some of the constituent gene arrays might meet each of these definitions. Taken as a whole, each neighbourhood represents the repertoire of alternative configurations of genes within a distinct set of genes, which form various (predicted) operons. Overlapping portions of these neighbourhoods are, to a varying extent, conserved during evolution, which confers functional relevance on the entire neighbourhoods.[33]

Genomic hitchhiking

Systematic analysis of evolutionarily conserved gene neighbourhoods showed that most of them consist predominantly of genes united by a coherent functional theme, but also include a minority of genes without an obvious functional connection to the main theme.[33] Although some of the latter genes might have unsuspected roles related to the main theme, others might be maintained within conserved gene arrays because of the advantage of expression at the level that is typical of the given neighbourhood. This phenomenon was designated 'genomic hitchhiking'.[33] In this study, the largest conserved neighbourhood included 79 genes (COGs) and consisted of overlapping, rearranged ribosomal protein superoperons; apparent genomic hitchhiking is particularly common in this neighbourhood and other neighbourhoods that consist of genes coding for translation machinery components (Figure 2).

Collinearity-free approach

SNAP algorithm

A new computational method, SNAP (similarity-neighbourhood approach), for finding functionally related gene sets from genomic context has been recently developed.[61,62] The novel feature of SNAP is that it does not rely on detection of conserved, collinear gene strings. Instead, a similarity-neighbourhood graph (SN-graph), which is constructed from the chains of similarity and neighbourhood relationships between

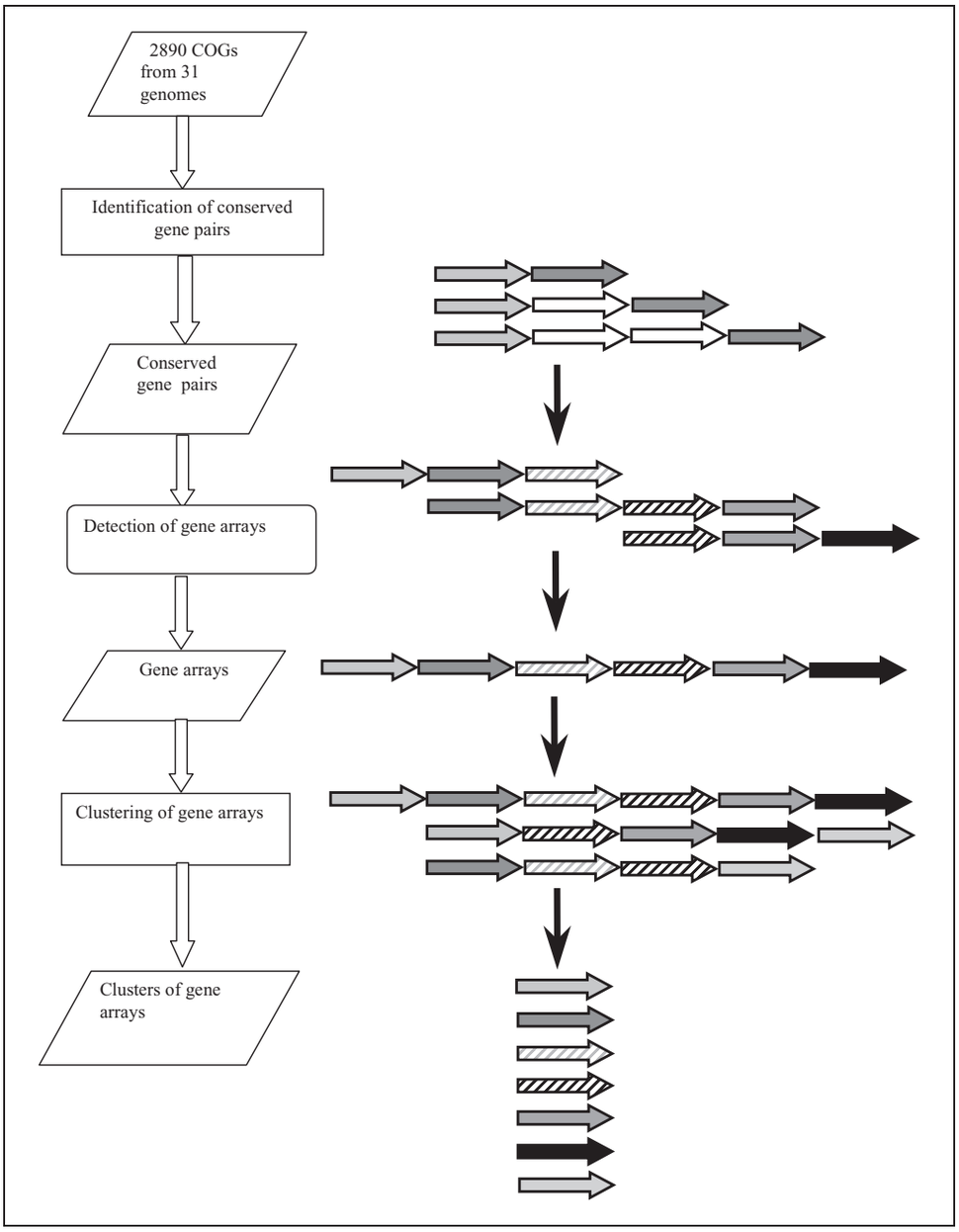

Figure 6: Flow chart of the procedure for construction of gene arrays and clusters from conserved gene pairs. Shaded or hatched arrows indicate COGs that form conserved pairs and empty arrows indicate COGs or non-COG genes that do not form conserved pairs, but are allowed to insert between genes in a conserved pair

Context methods for functional annotation

orthologous genes in different genomes and adjacent genes in the same genome, was introduced. An SN-cycle was defined as a closed path on the SN-graph. It has been demonstrated that SN-cycles derived from prokaryotic genome comparisons were substantially non-random and apparently functionally relevant. However, this approach is computationally intensive and is applicable to a limited number of genomes (< 30).[61]

IMPLICATIONS OF GENE ORDER CONSERVATION

Functional inferences

The most practically important consequence of gene order conservation is the possibility of functional prediction for uncharacterised genes. Comparative analysis of gene orders is one of the most powerful approaches among the so-called context methods of functional annotation of prokaryotic genomes.[14,63] We have already mentioned some examples of such

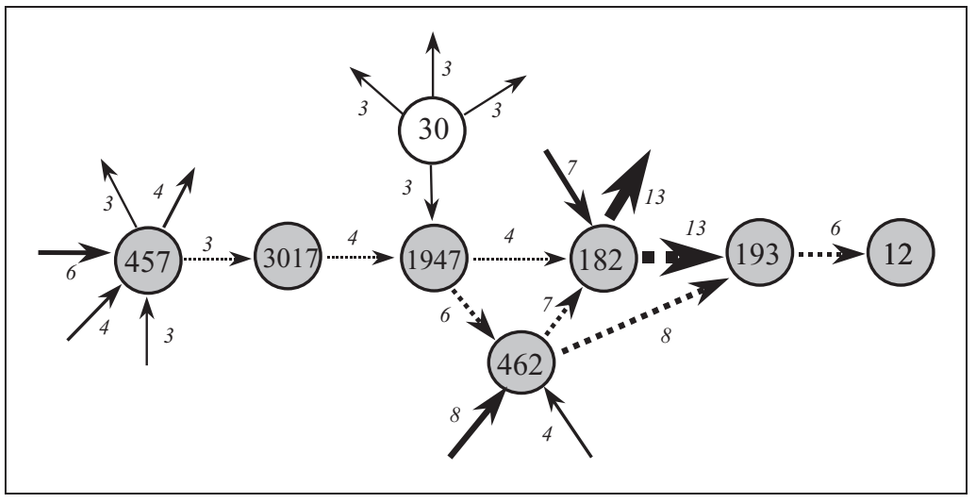

Figure 7: A cluster of gene arrays presented as an oriented graph. Nodes correspond to COGs, the COG numbers are indicated inside the circles. The edges show conserved gene pairs and the direction of transcription of the corresponding genes is shown by arrows. The grey circles and dotted arrows show the depicted cluster. The white circles and solid arrows show genes and gene pairs that are linked to individual COGs in the given cluster, but did not join it under the employed procedure. The number of genomes, in which the given pair is represented, is given for each edge, and the thickness of the edge is roughly proportional to this number. Definitions: COG0012, predicted GTPase (probably a translation factor); COG0193, peptidyl-tRNA hydrolase; COG0457, TPR-repeat-containing proteins; COG0462, phosphoribosylpyrophosphate synthetase; COG1207, *N*-acetylglucosamine-1-phosphate uridylyltransferase; COG1825, ribosomal protein L25; COG1947, 4-diphosphocytidyl-2C-methyl-D-erythritol 2-phosphate synthase; COG3017, outer membrane lipoprotein (outer membrane biogenesis)

predictions but it is worth briefly discussing additional cases yielded by more advanced methods for gene order comparison and/or by combination of gene order analysis and in-depth sequence comparison. For example, during a systematic analysis of conserved gene context in prokaryotic genomes,[33] a previously undetected, complex, partially conserved neighbourhood consisting of more than 20 genes was discovered in most Archaea and some bacteria, in particular, the hyperthermophiles *Thermotoga maritima* and *Aquifex aeolicus*.[64] The gene composition and gene order in this neighbourhood vary greatly among species, but all versions have a stable, conserved core that consists of five genes. One of the core genes encodes a predicted DNA helicase, often fused to a predicted HD-superfamily hydrolase (HD = histidine–aspartic acid), and another encodes a RecB family exonuclease; three core genes remain uncharacterised, but one of these might encode a nuclease of a new family. Two more genes that belong to this neighbourhood and are present in most of the genomes, in which the neighbourhood was detected, encode, respectively, another predicted HD-superfamily hydrolase (possibly a nuclease) of a distinct family and a predicted, novel DNA polymerase. The functional features of the proteins encoded in this neighbourhood suggest that they constitute a previously undetected DNA repair system, which is the first repair system largely specific for thermophiles to be identified. This hypothetical repair system might be functionally analogous to the bacterial-eukaryotic system of mutagenic translesion repair whose central components are DNA polymerases of the UmuC-DinB-Rad30-Rev1 superfamily, which typically are missing in thermophiles.[64]

New DNA repair systems

Missing enzymes in metabolic pathways

Another novel repair system in prokaryotes was predicted through a combination of detailed sequence analysis, phyletic profile examination and gene order comparison.[65] This gene complex consists primarily of genes that are generally associated with eukaryotic, rather than prokaryotic, replication and repair, namely, ATP-dependent DNA ligase, the small, catalytic subunit of the archaeo-eukaryotic-type DNA primase, and the homologues of the eukaryotic DNA-binding protein Ku (Figure 8). In several genomes, this neighbourhood also includes an uncharacterised bacterial-specific gene, possibly coding for a novel nuclease. On the basis of the function of the Ku protein in eukaryotes, it has been predicted that these putative operons might encode an uncharacterised complex involved in double-strand break repair.[65] This prediction was subsequently supported by demonstration of the cooperation between the mycobacterial homologue of Ku and ATP-dependent ligase encoded in the same neighbourhood in non-homologous DNA end joining.[66] The characterisation of the entire complex and elucidation of the role of the primase homologue await further experiments.

Recently, Osterman and Overbeek conceptualized a promising strategy for prediction of missing enzymes in metabolic pathways by using information on gene clustering and protein fusions combined with patterns of gene presence–absence in microbial species.[20] Missing links are identified by compiling evidence supporting the existence of a specific pathway in the analysed organism and revealing essential steps that cannot be connected to specific genes. Various techniques of genome context analysis[67] (Table 1) are used to infer functional coupling between genes coding for the known enzymes of the given pathways and uncharacterised genes; as a result, a

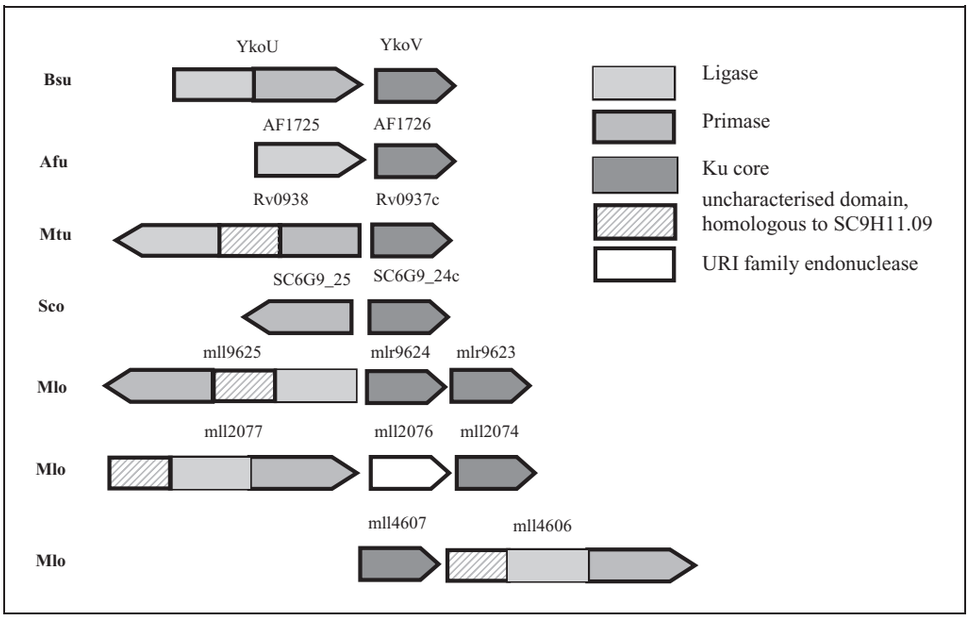

Figure 8: Organisation of genes and potential operons in the genomic regions coding for protein components of the predicted novel DNA repair system. Genes are shown not to scale; the direction of transcription is indicated by arrows. Orthologous genes are shown by the same pattern. Abbreviations: ligase, ATP-dependent DNA ligase; primase, the small, catalytic subunit of the archaeo-eukaryotic-type DNA primase; Ku core, the homologues of the eukaryotic DNA-binding protein Ku; Bsu *Bacillus subtilis*, Afu *Archaeoglobus fulgidus*, Mtu *Mycobacterium tuberculosis*, Sco *Streptomyces coelicolor*, Mlo *Mezorhizobium loti*

list of candidate genes for the missing function(s) is produced.[20] This approach was used for analysis of fatty acid biosynthesis in *Streptococcus pneumoniae*.[68] Almost all essential components of the fatty acid biosynthesis complex can be projected from *E. coli* to *S. pneumoniae* except for the *fabI* gene encoding enoyl-ACP-reductase. A novel *S. pneumoniae* enoyl-ACP-reductase (gene fabK) was predicted on the basis of gene order conservation, and this prediction was confirmed experimentally.[68]

Divergent overlapping genes may be artefacts of annotation

Functional signal analysis in prokaryotes

Gene order could be a valuable source of information for analysis of functional *cis*-signals in prokaryotic species. Spacers between convergent genes contain exclusively terminators, and spacers between divergent genes contain only promoters and signals that regulate transcription initiation, such as operators. Recognition of transcription regulatory sites in bacterial genomes is a hard problem. Generally, there are no algorithms capable of making robust predictions even for well-studied sites. However, availability of complete bacterial genomes allows one to increase the reliability of predictions by combining comparative analysis and gene order conservation.[69–71] This comparative approach is based on the assumption that sets of co-regulated genes are conserved in closely related bacteria. Thus, functionally relevant sites occur upstream of orthologous gene clusters, whereas false candidates are randomly scattered. This means not only that knowledge about regulation in well-studied genomes can be transferred to newly sequenced ones, but also that new members of regulons can be found.[69–71]

Combining functional signal prediction and gene order conservation

Analysis of overlapping genes

Overlapping coding regions in pro- and eukaryotic genomes are not necessarily artefacts, and some of them are evolutionarily conserved. It is well known that some unidirectional genes (Figure 3) overlap, typically by only a few nucleotides. Overlapping genes within operons might enable translational coupling[72] and/or protection of mRNA from degradation by maintaining its association with ribosomes.[73] Many of the overlapping convergent gene pairs (Figure 3) could represent real gene arrangements as suggested by comparative analysis.[74] In contrast, overlapping divergent coding regions (Figure 3) are more likely to be artefacts than overlapping convergent and unidirectional genes because a proper accommodation of promoter sequences within coding regions would be extremely hard to achieve. However, numerous pairs of divergent overlapping genes were detected (I. B. Rogozin, unpublished observations). It appears likely that most, if not all of them are artefacts caused by incorrect identification of the start of the involved coding regions.

Phylogenetic analysis of prokaryotes based on gene order conservation

Rearrangements continuously shuffle prokaryotic genomes, gradually breaking ancestral gene strings. Hence the natural idea to employ comparison of gene orders for phylogenetic reconstructions: in principle, the shorter the evolutionary distance between two genomes, the greater the extent of gene order conservation. The operonic organisation of a prokaryotic genome complicates the kinetics of this process. The apparent selective advantage of physical proximity for co-regulation makes some gene arrays less prone to break-up than others, thus extending the range of evolutionary distances over which gene order comparison is feasible.[59,75] However, selective forces acting on operons make them sensitive to the influence of the environmental niche occupied by the organism at different times during its evolutionary history. Furthermore, some operons appear to be particularly prone to being transferred as a whole, in accord with the selfish operon hypothesis,

Phylogenetic trees based on gene order comparison

accentuating the effect of horizontal transfer on the tree topology.[8]

Gene order comparisons yield predictions of gene function

To use gene order for phylogeny or similarity dendrogram construction, one has to choose a method for translating gene order data into a tree structure. This goal can be achieved by various means. Wolf and coworkers[76] employed the COG database to identify pairs of genes whose physical proximity was conserved in several genomes. The presence–absence matrices of these gene pairs were analysed using Dollo parsimony and neighbour joining methods, which produced essentially the same topology. Korbel and coworkers[31] counted adjacent pairs of BeT-derived orthologues shared by two genomes, converted the fraction of such pairs to distance and used these distances to construct neighbour-joining or least-squares trees. This technique is available at the SHOT web server (Table 1).[31] Owing to the high rate of intragenomic rearrangements, the gene order trees are, at least in theory, especially suitable to resolving the phylogeny of closely related prokaryotic species.[46] Generally, this approach behaves in a manner similar to the gene content methods (reviewed in Wolf *et al.*[75]), providing a good separation between the highest taxa, such as archaea and bacteria, and keeping closely related species together, but offering poor resolution on intermediate distances. Both groups described the effect of horizontal gene transfer on these trees.[75] Several additional approaches for phylogenetic reconstruction based on gene order comparison were recently reviewed by Sawa and coworkers.[77]

CONCLUSIONS

Because of the rapid evolution of gene order in prokaryotes, the potential of genome alignments for prediction of gene functions is limited. Nevertheless, such predictions yield valuable information, which is often distinct from and substantially complements results obtained through protein sequence and structure analysis. Furthermore, the utility of gene order analysis is further enhanced by the progress of genome sequencing as additional genomes increase the coverage of each individual genome with conserved gene strings. However, straightforward functional inferences from gene order conservation should be made with caution, given the relatively common instances of genomic hitchhiking. On the whole, applications of methods for gene order comparison yield a wealth of functional and evolutionary information that should be interpreted within the more general framework of genome context analysis and evolutionary conservation.[14,15,20,63]

The first tools for such integrated genomic context analysis have already been developed and can be used for rapid detection of potentially important links between genes (Table 1). This new generation of genome analysis tools includes STRING,[78] ERGO[67] and KEGG.[79] The challenge for the future is to develop robust criteria for combining different aspects of genomic context for reliable prediction of functional associations.

References

1. Dobzhansky, T. and Sturtevant, A. H. (1938), 'Inversions in the chromosomes of *Drosophila pseudoobscura*', *Genetics*, Vol. 23, pp. 28–64.
2. Jacob, F., Perrin, D., Sanchez, C. and Monod, J. (1960), 'L'Operon: Groupe de genes a expression coordonee par un operateur', *C. R. Seance Acad. Sci.*, Vol. 250, pp. 1727–1729.
3. Jacob, F. and Monod, J. (1961), 'Genetic regulatory mechanisms in the synthesis of proteins', *J. Mol. Biol.*, Vol. 3, pp. 318–356.
4. Miller, J. H. and Reznikoff, W. S. E. (1978), 'The Operon', Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
5. Mushegian, A. R. and Koonin, E. V. (1996), 'Gene order is not conserved in bacterial evolution', *Trends Genet.*, Vol. 12, pp. 289–290.
6. Watanabe, H., Mori, H., Itoh, T. and Gojobori, T. (1997), 'Genome plasticity as a paradigm of eubacteria evolution', *J. Mol. Evol.*, Vol. 44, Suppl 1, pp. S57–64.
7. Dandekar, T., Snel, B., Huynen, M. and Bork, P. (1998), 'Conservation of gene order: A

fingerprint of proteins that physically interact', *Trends Biochem. Sci.*, Vol. 23, pp. 324–328.

8. Lawrence, J. G. (1999), 'Selfish operons: The evolutionary impact of gene clustering in prokaryotes and eukaryotes', *Curr. Opin. Genet. Dev.*, Vol. 9, pp. 642–648.
9. Lawrence, J. G. (2002), 'Shared strategies in gene organization among prokaryotes and eukaryotes', *Cell*, Vol. 110, pp. 407–413.
10. Pevzner, P. and Tesler, G. (2003), 'Genome rearrangements in mammalian evolution: Lessons from human and mouse genomes', *Genome Res.*, Vol. 13, pp. 37–45.
11. Aravind, L. (2000), 'Guilt by association: Contextual information in genome analysis', *Genome Res.*, Vol. 10, pp. 1074–1077.
12. 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*, Vol. 402, pp. 86–90.
13. Marcotte, E. M., Pellegrini, M., Thompson, M. J. *et al.* (1999), 'A combined algorithm for genome-wide prediction of protein function', *Nature*, Vol. 402, pp. 83–86.
14. Huynen, M. A. and Snel, B. (2000), 'Gene and context: Integrative approaches to genome analysis', *Adv. Protein Chem.*, Vol. 54, pp. 345–379.
15. Galperin, M. Y. and Koonin, E. V. (2000), 'Who's your neighbor? New computational approaches for functional genomics', *Nature Biotechnol.*, Vol. 18, pp. 609–613.
16. Nierman, W. C., Eisen, J. A., Fleischmann, R. D. and Fraser, C. M. (2000), 'Genome data: What do we learn?' *Curr. Opin. Struct. Biol.*, Vol. 10, pp. 343–348.
17. Fujibuchi, W., Ogata, H., Matsuda, H. and Kanehisa, M. (2000), 'Automatic detection of conserved gene clusters in multiple genomes by graph comparison and P-quasi grouping', *Nucleic Acids Res.*, Vol. 28, pp. 4029–4036.
18. Zheng, Y., Szustakowski, J. D., Fortnow, L. *et al.* (2002), 'Computational identification of operons in microbial genomes', *Genome Res.*, Vol. 12, pp. 1221–1230.
19. Zheng, Y., Roberts, R .J. and Kasif, S. (2002), 'Genomic functional annotation using co-evolution profiles of gene clusters', *Genome Biol.*, Vol. 3, pp. R0060.1–60.9.
20. Osterman, A. and Overbeek, R. (2003), 'Missing genes in metabolic pathways: A comparative genomics approach', *Curr. Opin. Chem. Biol.*, Vol. 7, pp. 238–251.
21. 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*, Vol. 97, pp. 6652–6657.
22. Moreno-Hagelsieb, G. and Collado-Vides, J. (2002), 'A powerful non-homology method for the prediction of operons in prokaryotes', *Bioinformatics*, Vol. 18, Suppl 1, pp. S329–336.
23. Zuckerkandl, E. and Pauling, L. (1965), 'Molecules as documents of evolutionary history', *J. Theor. Biol.*, Vol. 8, pp. 357–366.
24. Fitch, W. M. (1970), 'Distinguishing homologous from analogous proteins', *Systematic Zoology*, Vol. 19, pp. 99–106.
25. Sonnhammer, E. L. and Koonin, E. V. (2002), 'Orthology, paralogy and proposed classification for paralog subtypes', *Trends Genet.*, Vol. 18, pp. 619–620.
26. Tatusov, R. L., Koonin, E. V. and Lipman, D. J. (1997), 'A genomic perspective on protein families', *Science*, Vol. 278, pp. 631–637.
27. Tatusov, R. L., Natale, D. A., Garkavtsev, I. V. *et al.* (2001), 'The COG database: New developments in phylogenetic classification of proteins from complete genomes', *Nucleic Acids Res.*, Vol. 29, pp. 22–28.
28. Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, B. *et al.* (2003), 'The COG database: An updated version includes eukaryotes', *BMC Bioinformatics*, Vol. 4, pp. 41.
29. Tatusov, R. L., Mushegian, A. R., Bork, P. *et al.* (1996), 'Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli*', *Curr. Biol.*, Vol. 6, pp. 279–291.
30. Wolf, Y. I., Rogozin, I. B., Kondrashov, A. S. and Koonin, E. V. (2001), 'Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context', *Genome Res.*, Vol. 11, pp. 356–372.
31. Korbel, J. O., Snel, B., Huynen, M. A. and Bork, P. (2002), 'SHOT: A web server for the construction of genome phylogenies', *Trends Genet.*, Vol. 18, pp. 158–162.
32. Lawrence, J. G. and Roth, J. R. (1996), 'Selfish operons: Horizontal transfer may drive the evolution of gene clusters', *Genetics*, Vol. 143, pp. 1843–1860.
33. Rogozin, I. B., Makarova, K. S., Murvai, J. *et al.* (2002), 'Connected gene neighborhoods in prokaryotic genomes', *Nucleic Acids Res*, Vol. 30, pp. 2212–2223.
34. Skovgaard, M., Jensen, L. J., Brunak, S. *et al.* (2001), 'On the total number of genes and their length distribution in complete microbial genomes', *Trends Genet.*, Vol. 17, pp. 425–428.
35. Devos, D. and Valencia, A. (2001), 'Intrinsic errors in genome annotation', *Trends Genet.*, Vol. 17, pp. 429–431.
36. Natale, D. A., Galperin, M. Y., Tatusov, R. L. and Koonin, E. V. (2000), 'Using the COG

database to improve gene recognition in complete genomes', *Genetica*, Vol. 108, pp. 9–17.

37. Brenner, S. E. (1999), 'Errors in genome annotation', *Trends Genet.*, Vol. 15, pp. 132–133.
38. Andersson, J. O. and Andersson, S. G. (2001), 'Pseudogenes, junk DNA, and the dynamics of *Rickettsia* genomes', *Mol. Biol. Evol.*, Vol. 18, pp. 829–839.
39. Ogata, H., Audic, S., Renesto-Audiffren, P. *et al.* (2001), 'Mechanisms of evolution in *Rickettsia conorii* and *R. prowazekii*', *Science*, Vol. 293, pp. 2093–2098.
40. Cole, S. T., Eiglmeier, K., Parkhill, J. *et al.* (2001), 'Massive gene decay in the leprosy bacillus', *Nature*, Vol. 409, pp. 1007–1011.
41. Needleman, S. B. and Wunsch, C. D. (1970), 'A general method applicable to the search for similarities in the amino acid sequence of two proteins', *J. Mol. Biol.*, Vol. 48, pp. 443–453.
42. Smith, T. F. and Waterman, M. S. (1981), 'Identification of common molecular subsequences', *J. Mol. Biol.*, Vol. 147, pp. 195–197.
43. Staden, R. (1977), 'Sequence data handling by computer', *Nucleic Acids Res.*, Vol. 4, pp. 4037–4051.
44. Tillier, E. R. and Collins, R. A. (2000), 'Genome rearrangement by replication-directed translocation', *Nat. Genet.*, Vol. 26, pp. 195–197.
45. Eisen, J. A., Heidelberg, J. F., White, O. and Salzberg, S. L. (2000), 'Evidence for symmetric chromosomal inversions around the replication origin in bacteria', *Genome Biol.*, Vol. 1, pp. R0011.1–11.9.
46. Suyama, M. and Bork, P. (2001), 'Evolution of prokaryotic gene order: Genome rearrangements in closely related species', *Trends Genet.*, Vol. 17, pp. 10–13.
47. 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.*, Vol. 16, pp. 332–346.
48. Ogata, H., Goto, S., Sato, K. *et al.* (1999), 'KEGG: Kyoto Encyclopedia of Genes and Genomes', *Nucleic Acids Res.*, Vol. 27, pp. 29–34.
49. Ogata, H., Fujibuchi, W., Goto, S. and Kanehisa, M. (2000), 'A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters', *Nucleic Acids Res.*, Vol. 28, pp. 4021–4028.
50. Koonin, E. V., Wolf, Y. I. and Aravind, L. (2001), 'Prediction of the archaeal exosome and its connections with the proteasome and the translation and transcription machineries by a comparative-genomic approach', *Genome Res.*, Vol. 11, pp. 240–252.
51. Evguenieva-Hackenburg, E., Walter, P., Hochleitner, E. *et al.* (2003), 'An exosome-like complex in *Sulfolobus solfataricus*', *EMBO Rep.*, Vol. 4, pp. 889–893.
52. Snel, B., Lehmann, G., Bork, P. and Huynen, M. A. (2000), 'STRING: A web-server to retrieve and display the repeatedly occurring neighbourhood of a gene', *Nucleic Acids Res.*, Vol. 28, pp. 3442–3444.
53. Overbeek, R., Fonstein, M., D'Souza, M. *et al.* (1999), 'The use of gene clusters to infer functional coupling', *Proc. Natl Acad. Sci. USA*, Vol. 96, pp. 2896–2901.
54. Ermolaeva, M. D., White, O. and Salzberg, S. L. (2001), 'Prediction of operons in microbial genomes', *Nucleic Acids Res.*, Vol. 29, pp. 1216–1221.
55. Yanai, I., Mellor, J. C. and DeLisi, C. (2002), 'Identifying functional links between genes using conserved chromosomal proximity', *Trends Genet.*, Vol. 18, pp. 176–179.
56. Tamames, J., Casari, G., Ouzounis, C. and Valencia, A. (1997), 'Conserved clusters of functionally related genes in two bacterial genomes', *J. Mol. Evol.*, Vol. 44, pp. 66–73.
57. Rogozin, I. B., Makarova, K. S., Natale, D. A. *et al.* (2002), 'Congruent evolution of different classes of non-coding DNA in prokaryotic genomes', *Nucleic Acids Res.*, Vol. 30, pp. 4264–4271.
58. Salgado, H., Santos-Zavaleta, A., Gama-Castro, S. *et al.* (2001), 'RegulonDB (version 3.2): transcriptional regulation and operon organization in *Escherichia coli* K-12', *Nucleic Acids Res.*, Vol. 29, pp. 72–74.
59. Lathe, W. C. 3rd, Snel, B. and Bork, P. (2000), 'Gene context conservation of a higher order than operons', *Trends Biochem Sci.*, Vol. 25, pp. 474–479.
60. Snel, B., Bork, P. and Huynen, M. (2002), 'Conservation of gene co-regulation in prokaryotes and eukaryotes', *Trends Biotechnol.*, Vol. 20, p. 410.
61. Kolesov, G., Mewes, H. W. and Frishman, D. (2001), 'SNAPping up functionally related genes based on context information: A colinearity-free approach', *J. Mol. Biol.*, Vol. 311, pp. 639–656.
62. Kolesov, G., Mewes, H. W. and Frishman, D. (2002), 'SNAPper: Gene order predicts gene function', *Bioinformatics*, Vol. 18, pp. 1017–1019.
63. Huynen, M., Snel, B., Lathe, W. and Bork, P. (2000), 'Exploitation of gene context', *Curr. Opin. Struct. Biol.*, Vol. 10, pp. 366–370.
64. Makarova, K. S., Aravind, L., Grishin, N. V. *et al.* (2002), 'A DNA repair system specific for thermophilic Archaea and bacteria predicted

by genomic context analysis', *Nucleic Acids Res.*, Vol. 30, pp. 482–496.

65. Aravind, L. and Koonin, E. V. (2001), 'Prokaryotic homologs of the eukaryotic DNA-end-binding protein Ku, novel domains in the Ku protein and prediction of a prokaryotic double-strand break repair system', *Genome Res.*, Vol. 11, pp. 1365–1374.
66. Weller, G. R., Kysela, B., Roy, R. *et al.* (2002), 'Identification of a DNA nonhomologous end-joining complex in bacteria', *Science*, Vol. 297, pp. 1686–1689.
67. Overbeek, R., Larsen, N., Walunas, T. *et al.* (2003), 'The ERGO genome analysis and discovery system', *Nucleic Acids Res.*, Vol. 31, pp. 164–171.
68. Heath, R. J. and Rock, C. O. (2000), 'A triclosan-resistant bacterial enzyme', *Nature*, Vol. 406, pp. 145–146.
69. Mironov, A. A., Koonin, E. V., Roytberg, M. A. and Gelfand, M. S. (1999), 'Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes', *Nucleic Acids Res.*, Vol. 27, pp. 2981–2989.
70. Gelfand, M. S., Novichkov, P. S., Novichkova, E. S. and Mironov, A. A. (2000), 'Comparative analysis of regulatory patterns in bacterial genomes', *Brief. Bioinform.*, Vol. 1, pp. 357–371.
71. Tan, K., Moreno-Hagelsieb, G., Collado-Vides, J. and Stormo, G. D. (2001), 'A comparative genomics approach to prediction of new members of regulons', *Genome Res.*, Vol. 11, pp. 566–584.
72. Oppenheim, D. S. and Yanofsky, C. (1980), 'Functional analysis of wild-type and altered tryptophan operon promoters of *Salmonella typhimurium* in *Escherichia coli*', *J. Mol. Biol.*, Vol. 144, pp. 143–161.
73. Schneider, E., Blundell, M. and Kennell, D. (1978), 'Translation and mRNA decay', *Mol. Gen. Genet.*, Vol. 160, pp. 121–129.
74. Rogozin, I. B., Spiridonov, A. M., Sorokin, A. V. *et al.* (2002), 'Purifying and directional selection in overlapping prokaryotic genes', *Trends Genet.*, Vol. 18, pp. 228–232.
75. Wolf, Y. I., Rogozin, I. B., Grishin, N. V. and Koonin, E. V. (2002), 'Genome trees and the tree of life', *Trends Genet.*, Vol. 18, pp. 472–479.
76. Wolf, Y. I., Rogozin, I. B., Grishin, N. V. *et al.* (2001), 'Genome trees constructed using five different approaches suggest new major bacterial clades', *BMC Evol. Biol.*, Vol. 1, p. 8.
77. Sawa, G., Dicks, J. and Roberts, I. N. (2003), 'Current approaches to whole genome phylogenetic analysis', *Brief. Bioinform.*, Vol. 4, pp. 63–74.
78. von Mering, C., Huynen, M., Jaeggi, D. *et al.* (2003), 'STRING: A database of predicted functional associations between proteins', *Nucleic Acids Res.*, Vol. 31, pp. 258–261.
79. Kanehisa, M., Goto, S., Kawashima, S. *et al.* (2004), 'The KEGG resource for deciphering the genome', *Nucleic Acids Res.*, Vol. 32, pp. D277–D280.