W176–W180 *Nucleic Acids Research, 2008, Vol. 36, Web Server issue* *Published online 29 May 2008*
doi:10.1093/nar/gkn330

GeConT 2: gene context analysis for orthologous proteins, conserved domains and metabolic pathways

C. E. Martinez-Guerrero[1], R. Ciria[1], C. Abreu-Goodger[2], G. Moreno-Hagelsieb[3] and E. Merino[1,*]

[1]Departmento de Microbiología Molecular, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, Morelos 62210, México, [2]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK and [3]Department of Biology, Wilfrid Laurier University, Waterloo, ON, N2L 3C5, Canada

Received January 31, 2008; Revised April 26, 2008; Accepted May 8, 2008

ABSTRACT

The Gene Context Tool (GeConT) allows users to visualize the genomic context of a gene or a group of genes and their orthologous relationships within fully sequenced bacterial genomes. The new version of the server incorporates information from the COG, Pfam and KEGG databases, allowing users to have an integrated graphical representation of the function of genes at multiple levels, their phylogenetic distribution and their genomic context. The sequence of any of the genes can be easily retrieved, as well as the 5′ or 3′ regulatory regions, greatly facilitating further types of analysis. GeConT 2 is available at: http://bioinfo.ibt.unam.mx/gecont.

INTRODUCTION

With more than 660 prokaryotic genomes in the current RefSeq database (1), the need for tools that allow biologists to visualize the genomic context of genes of their interest becomes crucial. Despite the tendency of prokaryotes towards very little overall conservation of gene order (2,3), groups of genes participating in particular functions tend to remain close together across different lineages, either as part of operons (4,5) or as functional neighbourhoods comprising several transcription units (6,7). Genomic context is not limited to genes close to each other. Overall, inference of functional associations from genomic context can be derived from the following kinds of evidence: (i) Gene fusions (8,9), whereby separated genes would be assumed to work together if they are found as a single, fused, gene in another organism; (ii) Conservation of gene order (3,10), where conservation across evolutionarily distant organisms is taken as evidence of functional association and (iii) similarity of phylogenetic profiles (9,11,12), whereby two genes are assumed to work together if their orthologs tend to co-occur, appear and disappear in concert, across different organisms, with the idea that genes working together would both either be present or absent because the presence of a single one of them would be useless without the other. A fourth evidence of functional interactions would be provided by the study of the rearrangement of operons across lineages (13–15). The idea here is that the rearrangements or reorganization of transcription units across genomes might be conservative in the sense that newly formed operons will put genes with related functions together, thus revealing a functional association that would not be apparent in a single organism.

Biologists interested in particular groups of genes or functional modules would be able to find other features by visually inspecting the genomic context or neighbourhood of the genes of their interest. Such experts might be able to interpret these neighbourhoods and find examples of non-orthologous gene displacement (2), or horizontal gene transfers, that might have an effect on the functional module in particular organisms. Further tests of the validity of their findings would be greatly facilitated if the tool used to visualize the gene modules across several genomes would also allow for downloading of meaningful sequences, such as the protein sequences, or the DNA coding for the gene, or the DNA sequences occurring downstream or upstream the gene. There are excellent tools that allow for the retrieval of functional predictions based on genomic context, such as STRING (16) and PROLINKS (17), but such tools restrict the visualization of genomic context to the particular predictions associated to a gene or genes of interest, rather than to any physical neighbours. Also, while protein sequences of predicted interactors can be retrieved from these servers, nucleotide sequences of genes or intergenic regions are not available.

*To whom correspondence should be addressed. Tel: +52 777 3291648; Fax: +52 777 3172388; Email: merino@ibt.unam.mx

In other instances, such as GECO (18), the navigation interface for retrieving this type of information is not simple and the orthology definition is different from any of the most commonly accepted standards. The SEED (19) is a fully automated web resource that analyses the genome context of bacterial and archaeal organisms, however its main purpose is oriented to genome annotation rather than the genome context exploration by a particular user who wants to examine the neighbourhoods of his/her genes of interest. Another useful web server is the comprehensive microbial resource (http://cmr.jcvi.org/tigr-scripts/CMR/CmrHomePage.cgi). Although this web server offers a wide variety of tools and resources to highlight differences and similarities between prokaryotic genomes, its comprehensiveness hinders straightforward navigation, further justifying the development of more simplified gene context analysis tools. Here we present Gene Context Tool (GeConT) 2, the second version of GeConT (20), a web-based tool that allows users to visualize their genes of interest, and their genomic context, across all available fully sequenced bacterial genomes. Orthologous domains are highlighted using shared colours. This makes it easy to navigate across the functional neighbourhood of any particular gene and its orthologs.

IMPROVEMENTS

GeConT 2 extends over the previous version in many ways. We have increased the query options to allow one or more of the following: (i) gene ids which can be given as common names, GI numbers as defined in GenBank (21) or SwissProt identifiers (22); (ii) orthologous groups as defined in the COG database (23); (iii) metabolic pathways as described in the KEGG database (24); (iv) protein domains taken from the Pfam database (25); (v) a protein or DNA sequence, from which similarities will be identified using the integrated BLAST (26) search; (vi) complex phrases, using Boolean operators, to allow flexible searches against all the descriptions of the included databases. Since many queries will likely result in hundreds of matches (many of which are likely to be redundant), we implemented a filter that can reduce the display to a user-specified number of non-redundant genomes. This option uses distances calculated from 16S rRNA alignments to select a set of representative genomes specific for each query. Additionally, the user can also restrict the search to particular phylogenetic groups of interest. The genome context can be displayed considering a user-defined number of flanking genes or in accordance to their predicted operon structure (27–29).

In agreement with the increased input flexibility, the output allows visualizing the genes colour-coded according to their COG, Pfam or KEGG assignations. Also new to this version, multiple domains can be visualized as distinct coloured regions within a gene. Domains, genes and intergenic regions are drawn to scale; overlapping genes and non-coding RNA genes are now included. The user can click on any cistron to display relevant information, including descriptions from COG, KEGG and Pfam, as well as the amino acid and nucleotide sequence, and upstream and downstream intergenic flanking regions. GeConT 2 also allows the user to retrieve all the sequence data of the set of genes that have been matched by the input query, facilitating further analysis.

WEB SERVER DESCRIPTION

All the code for GeConT 2 is written in Perl, generating HTML and JavaScript code on the fly, using the GD library for the dynamic creation of images. The server uses fully sequenced genomes downloaded from GenBank. All gene coordinates, DNA strand, names and descriptions are taken from these files. Pfam and COG annotations were computed for all coding sequences using the HMMER package (30). Pfam-A models (25) were directly obtained from ftp://ftp.sanger.ac.uk/pub/databases/Pfam/. COG models were generated by aligning the sequences from every COG with MUSCLE (31) and building the models with hmmerbuild from the HMMER package. KEGG pathway annotations for all genomes were downloaded from ftp://ftp.genome.jp/pub/kegg/pathway/. The resulting annotations for each gene are saved as indexed files that are tied to hashed arrays for faster access. When queried for a particular gene, the server calculates the sizes, distances and neighbours based on the stored information. Once the list of genes to be displayed has been calculated, the server assigns colours starting with the COG, Pfam or KEGG most represented among these genes. In this way, the user can gain a visual insight of the most abundant annotations among the displayed genes. Additionally, the information about any gene can be quickly inspected by placing the mouse over it.

EXAMPLES AND DISCUSSION

With GeConT 2 users will be able to perform fast, integrated and intuitive analyses in fully sequenced genomes. In this section we discuss several examples that help illustrate the functionality of the webserver.

Identifying conserved elements involved in regulating a given pathway

An important feature of GeConT 2 is its potential to do comparative genome analysis of related genes to look for potential conserved regulatory motifs. The gene relationship can be established based on their orthology or biochemical pathway associations as defined in the COG, KEGG and Pfam databases. Since regulatory elements are commonly more conserved in closely related organisms, users can restrict their searches to a particular phylogenetic group. For example, in order to identify likely regulatory elements in methionine metabolism represented by the KEEG pathway 00271 in Firmicutes and Proteobacteria, the user can perform the corresponding searches in these groups by using the 'Specific taxonomy' option. The output of two representative organisms of these groups is shown in Figure 1 of the Supplementary Material. Using the operon clustering option, and colouring by COG attributes, there are 17 different operons with enzymes related to methionine metabolism in the

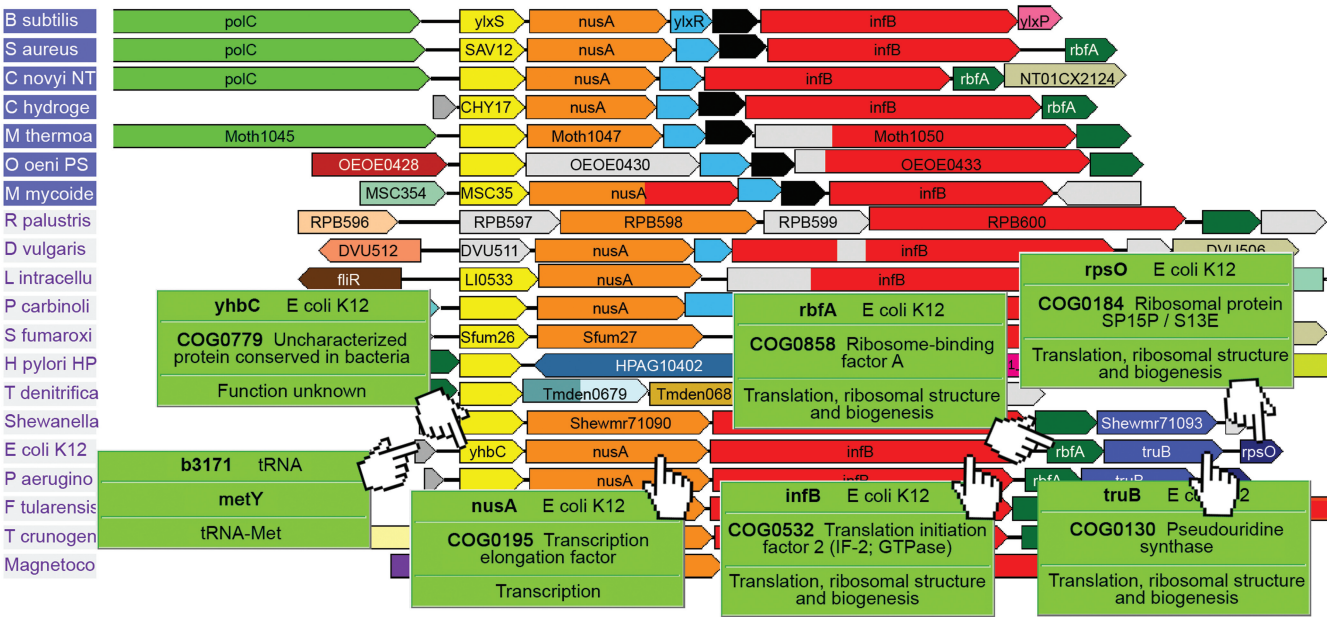

Figure 1. Genome context of COG0779/Pfam Duf150. Genome context analysis shows that all the conserved neighbouring genes of COG0779 have functions related to translation.

Firmicute *Bacillus halodurans*, while there are 11 operons in the Proteobacteria *Caulobacter crescentus*. The user can take advantage of the sequence retrieval options in GeConT 2 and get all the 5′ upstream regions for these operons. Using these sequences as the input of motif discovery programs such as MEME (32), the user can verify that the operons involved in this pathway are regulated by the SAM-I and S(MK) riboswitches in Firmicutes and by SAM-II in alpha-Proteobacteria (33–35). It is important to note that redundant information coming from different strains of the same organism might generate over-representation of particular sequences in the data set. To overcome this problem, the user can reduce the number of organisms returned by the 'maximum genomes to display' option. Previously we have shown the power of this kind of approach for identifying riboswitches starting from the regulatory regions of genes belonging to a same COG (36). It is now possible, using only web-based tools such as GeConT 2, to perform similar searches using any group of genes or pathways that a user might be interested in.

Functional insights for genes of unknown function

Most genes have little or no functional annotation. Even for the most studied bacteria, *Escherichia coli*, the fraction of genes for which detailed knowledge is available is still low [54% in the latest survey (37)]. Genomic context can give valuable insights into the functional relationships between neighbouring genes, for reasons discussed in the introduction. There are many cases of conserved proteins for which no functional assignation is available in the public databases. Homology searches are of no use, since all the hits also lack function. Context analysis can help solve some of these cases. One such example is annotated in Pfam as Duf150 (Domain of Unknown Function 150) and in the COG database as COG0779 ('Uncharacterized protein conserved in bacteria'). Figure 1 shows a section of the results when searching for Duf150 in GeConT 2. It is easy to see that the context of this protein is well conserved, and the mouse-over function allows a quick view of the functional assignations of the neighbouring genes, all of which seem to be involved in transcriptional elongation or translation initiation. It is thus quite likely that Duf150/COG0779 members are functionally related to these processes, and this can be considered as a first general function prediction for these previously uncharacterized proteins.

Using context to discover the correct function for paralogs

When multiple copies of a gene arise by duplication (paralogs), it can become particularly difficult to assign the correct function, at least by sequence alone. For example, the enzymes TrpE (Anthranilate synthases component I) and PabB (para-aminobenzoate synthases component I) have great sequence similarity and perform very similar reactions. These enzymes use pyruvate as a common substrate, although they participate in different pathways involved in tryptophan and in folate biosynthesis, respectively. Based on genome context, the *trpE* and *pabB* genes can easily be distinguished even in un-annotated genomes. With GeConT 2 we can analyse the neighbourhood searching for the other genes of the corresponding metabolic pathways (Figure 2). Another good example of paralogous domains is Palp (Pyridoxal-phosphate-dependant enzyme). Enzymes with this Pfam domain are highly versatile, participating in the biosynthesis of different amino acids such as tryptophan, cysteine, serine and threonine. Again, the context as well as the COG annotations allow us to easily distinguish between the different

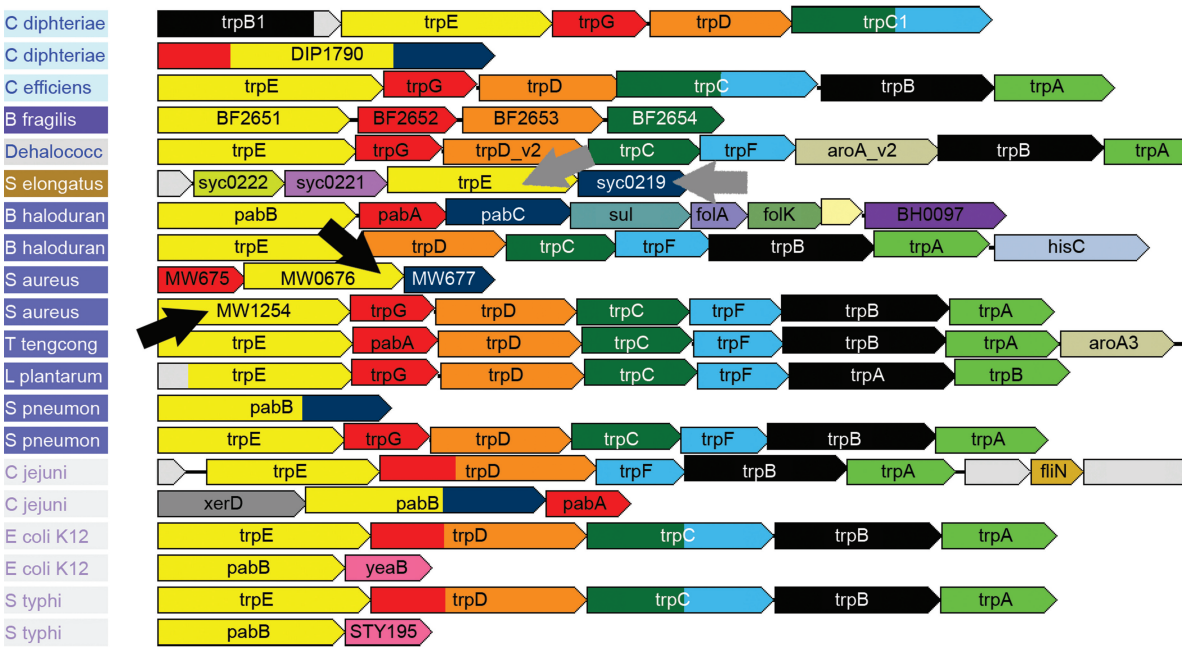

Figure 2. Identification of gene function among paralogous genes, based on the operon structure and COG annotations. The *trpE* genes can be differentiated from *pabB* (both in yellow) since the first one is transcribed with other genes of the tryptophan biosynthetic pathway, such as *trpD* (orange) *trpC* (dark green), *trpF* (light blue), while *pabB* is part of operons carrying genes of the folate biosynthetic pathway, such as *pabC* (dark blue). In *Staphylococcus aureus* MW2 *trpE* and *pabB* are not annotated (black arrows), yet we can clearly distinguish them from their context. We can also see that in *Synechococcus elongatus*, *trpE* is incorrectly annotated since this gene is co-transcribed with *pabC* (grey arrows).

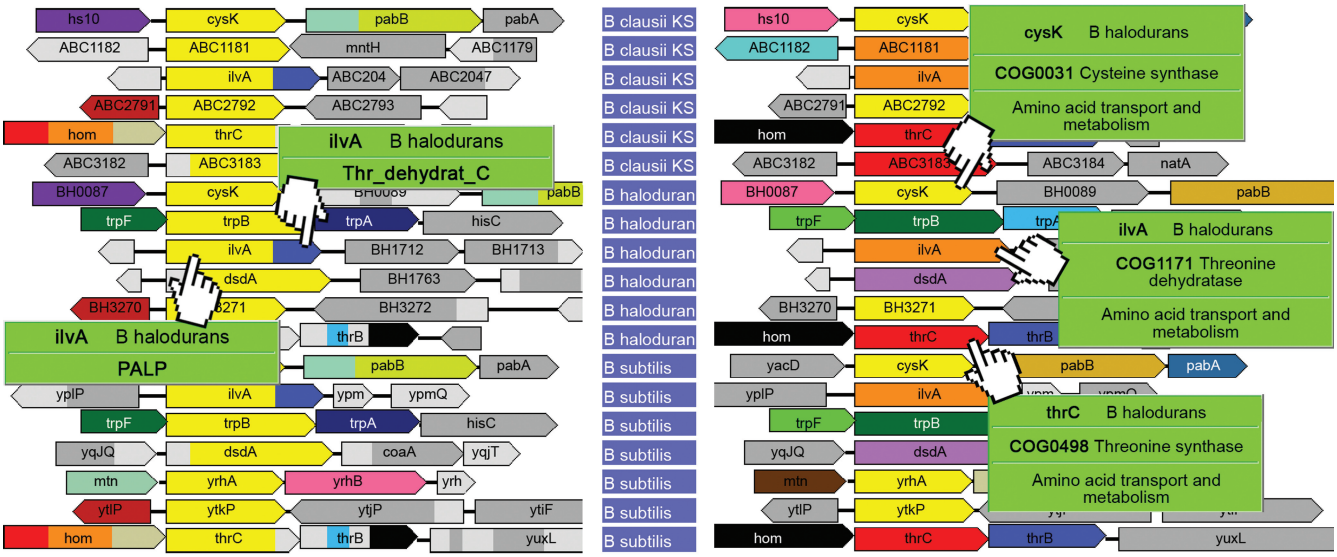

Figure 3. Genome context of the Pfam Palp domain in Firmicutes. One of the most salient features of GeConT 2 is the possibility of displaying contexts of multiple instances of a domain within a single genome, besides corresponding contexts in other genomes. On the left, genes are coloured by Pfam domains, with the yellow one corresponding to the Pfam Palp domain. On the right, the same genes are coloured by COG. This analysis shows how this catalytic domain can be used for different purposes by enzymes involved in the biosynthesis of different amino acids.

pathways and correctly identify the specific function of each gene (Figure 3).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We wish to thank Shirley Ainsworth for bibliographical assistance and Abel Linares, Arturo Ocadiz, Juan Manuel Hurtado, Alma Martinez and Nancy Mena for computer support. G.M-H. acknowledges computer facilities of the Shared Hierarchical Academic Research Computing

Network (SHARCNET). Funding was provided by Natural Sciences and Engineering Research Council of Canada (NSERC) to G.M-H. Sanger Institute Postdoctoral Fellowship to C.A-G. Consejo Nacional de Ciencia y Tecnología (CONACyT) [60127-Q] and PAPIIT IN212708 grants to E.M. Macroproyecto de Tecnologías de la información y la computación-UNAM to E.M. Funding to pay the Open Access publication charges for this article was provided by CONACYT [60127-Q].

Conflict of interest statement. None declared.

REFERENCES

1. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
2. Koonin,E.V., Mushegian,A.R. and Bork,P. (1996) Non-orthologous gene displacement. *Trends Genet.*, **12**, 334–336.
3. Bork,P., Dandekar,T., Diaz-Lazcoz,Y., Eisenhaber,F., Huynen,M. and Yuan,Y. (1998) Predicting function: from genes to genomes and back. *J. Mol. Biol.*, **283**, 707–725.
4. Ermolaeva,M.D., White,O. and Salzberg,S.L. (2001) Prediction of operons in microbial genomes. *Nucleic Acids Res.*, **29**, 1216–1221.
5. Moreno-Hagelsieb,G., Trevino,V., Perez-Rueda,E., Smith,T.F. and Collado-Vides,J. (2001) Transcription unit conservation in the three domains of life: a perspective from Escherichia coli. *Trends Genet.*, **17**, 175–177.
6. Tamames,J., Casari,G., Ouzounis,C. and Valencia,A. (1997) Conserved clusters of functionally related genes in two bacterial genomes. *J. Mol. Evol.*, **44**, 66–73.
7. Galperin,M.Y. and Koonin,E.V. (2000) Who's your neighbor? New computational approaches for functional genomics. *Nat. Biotechnol.*, **18**, 609–613.
8. Enright,A.J., Iliopoulos,I., Kyrpides,N.C. and Ouzounis,C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.
9. Marcotte,E.M., Pellegrini,M., Ng,H.L., Rice,D.W., Yeates,T.O. and Eisenberg,D. (1999) Detecting protein function and protein-protein interactions from genome sequences. *Science*, **285**, 751–753.
10. Overbeek,R., Fonstein,M., D'Souza,M., Pusch,G.D. and Maltsev,N. (1999) The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA*, **96**, 2896–2901.
11. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
12. Gaasterland,T. and Ragan,M.A. (1998) Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes. *Microb. Comp. Genomics*, **3**, 199–217.
13. Rogozin,I.B., Makarova,K.S., Murvai,J., Czabarka,E., Wolf,Y.I., Tatusov,R.L., Szekely,L.A. and Koonin,E.V. (2002) Connected gene neighborhoods in prokaryotic genomes. *Nucleic Acids Res.*, **30**, 2212–2223.
14. Snel,B., Bork,P. and Huynen,M.A. (2002) The identification of functional modules from the genomic association of genes. *Proc. Natl Acad. Sci. USA*, **99**, 5890–5895.
15. Janga,S.C., Collado-Vides,J. and Moreno-Hagelsieb,G. (2005) Nebulon: a system for the inference of functional relationships of gene products from the rearrangement of predicted operons. *Nucleic Acids Res.*, **33**, 2521–2530.
16. von Mering,C., Jensen,L.J., Kuhn,M., Chaffron,S., Doerks,T., Kruger,B., Snel,B. and Bork,P. (2007) STRING 7—recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.*, **35**, D358–D362.
17. Bowers,P.M., Pellegrini,M., Thompson,M.J., Fierro,J., Yeates,T.O. and Eisenberg,D. (2004) Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol.*, **5**, R35.
18. Kuenne,C.T., Ghai,R., Chakraborty,T. and Hain,T. (2007) GECO—linear visualization for comparative genomics. *Bioinformatics*, **23**, 125–126.
19. Overbeek,R., Begley,T., Butler,R.M., Choudhuri,J.V., Chuang,H.Y., Cohoon,M., de Crecy-Lagard,V., Diaz,N., Disz,T., Edwards,R. *et al.* (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.*, **33**, 5691–5702.
20. Ciria,R., Abreu-Goodger,C., Morett,E. and Merino,E. (2004) GeConT: gene context analysis. *Bioinformatics*, **20**, 2307–2308.
21. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2008) GenBank. *Nucleic Acids Res.*, **36**, D25–D30.
22. Consortium,T.U. (2008) The universal protein resource (UniProt). *Nucleic Acids Res.*, **36**, D190–D195.
23. Tatusov,R.L., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Kiryutin,B., Koonin,E.V., Krylov,D.M., Mazumder,R., Mekhedov,S.L., Nikolskaya,A.N. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
24. Kanehisa,M., Araki,M., Goto,S., Hattori,M., Hirakawa,M., Itoh,M., Katayama,T., Kawashima,S., Okuda,S., Tokimatsu,T. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.
25. Finn,R.D., Tate,J., Mistry,J., Coggill,P.C., Sammut,S.J., Hotz,H.R., Ceric,G., Forslund,K., Eddy,S.R., Sonnhammer,E.L. *et al.* (2008) The Pfam protein families database. *Nucleic Acids Res.*, **36**, D281–D288.
26. Schaffer,A.A., Aravind,L., Madden,T.L., Shavirin,S., Spouge,J.L., Wolf,Y.I., Koonin,E.V. and Altschul,S.F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.*, **29**, 2994–3005.
27. Moreno-Hagelsieb,G. and Collado-Vides,J. (2002) A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics*, **18 (Suppl. 1)**, S329–S336.
28. Janga,S.C. and Moreno-Hagelsieb,G. (2004) Conservation of adjacency as evidence of paralogous operons. *Nucleic Acids Res.*, **32**, 5392–5397.
29. Moreno-Hagelsieb,G. (2006) Operons across prokaryotes: genomic analyses and predictions 300+ genomes later. *Curr. Genomics*, **7**, 163–170.
30. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
31. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
32. Bailey,T.L., Williams,N., Misleh,C. and Li,W.W. (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.*, **34**, W369–W373.
33. Corbino,K.A., Barrick,J.E., Lim,J., Welz,R., Tucker,B.J., Puskarz,I., Mandal,M., Rudnick,N.D. and Breaker,R.R. (2005) Evidence for a second class of S-adenosylmethionine riboswitches and other regulatory RNA motifs in alpha-proteobacteria. *Genome Biol.*, **6**, R70.
34. Fuchs,R.T., Grundy,F.J. and Henkin,T.M. (2006) The S(MK) box is a new SAM-binding RNA for translational regulation of SAM synthetase. *Nat. Struct. Mol. Biol.*, **13**, 226–233.
35. Grundy,F.J. and Henkin,T.M. (1998) The S box regulon: a new global transcription termination control system for methionine and cysteine biosynthesis genes in gram-positive bacteria. *Mol. Microbiol.*, **30**, 737–749.
36. Abreu-Goodger,C., Ontiveros-Palacios,N., Ciria,R. and Merino,E. (2004) Conserved regulatory motifs in bacteria: riboswitches and beyond. *Trends Genet.*, **20**, 475–479.
37. Riley,M., Abe,T., Arnaud,M.B., Berlyn,M.K., Blattner,F.R., Chaudhuri,R.R., Glasner,J.D., Horiuchi,T., Keseler,I.M., Kosuge,T. *et al.* (2006) Escherichia coli K-12: a cooperatively developed annotation snapshot—2005. *Nucleic Acids Res.*, **34**, 1–9.