Published online 23 November 2009 *Nucleic Acids Research, 2010, Vol. 38, Database issue* ***D57–D61***
doi:10.1093/nar/gkp938

ACLAME: A CLAssification of Mobile genetic Elements, update 2010

Raphaël Leplae*, Gipsi Lima-Mendez and Ariane Toussaint

Bioinformatique des Génomes et des Réseaux, Université Libre de Bruxelles, Boulevard du Triomphe, 1050 Bruxelles, Belgium

Received September 15, 2009; Revised and Accepted October 9, 2009

ABSTRACT

The ACLAME database is dedicated to the collection, analysis and classification of sequenced mobile genetic elements (MGEs, in particular phages and plasmids). In addition to providing information on the MGEs content, classifications are available at various levels of organization. At the gene/protein level, families group similar sequences that are expected to share the same function. Families of four or more proteins are manually assigned with a functional annotation using the GeneOntology and the locally developed ontology MeGO dedicated to MGEs. At the genome level, evolutionary cohesive modules group sets of protein families shared among MGEs. At the population level, networks display the reticulate evolutionary relationships among MGEs. To increase the coverage of the phage sequence space, ACLAME version 0.4 incorporates 760 high-quality predicted prophages selected from the Prophinder database. Most of the data can be downloaded from the freely accessible ACLAME web site (http://aclame.ulb.ac.be). The BLAST interface for querying the database has been extended and numerous tools for in-depth analysis of the results have been added.

INTRODUCTION

With the advent of new sequencing techniques, the amount of fully sequenced bacterial genomes is increasing at an unprecedented rate. In addition, metagenomic projects are providing massive amount of genomic sequence fragments for which bioinformatics analysis remains a challenging task both technically and scientifically. However, a common observation derived so far from large-scale (meta)genomic sequence analysis is the complexity of bacterial genomes, at the structural and compositional levels. The major contributors to this 'unexpected' diversity are the mobile genetic elements (MGEs) responsible for the widespread lateral gene transfers and genomes rearrangements.

Phages, plasmids, transposons and insertion sequences (IS) are members of the MGE population [generally called the *mobilome* (1)], a central player in mobilizing and reorganizing genes, be it within a given genome (intracellular mobility) or between bacterial cells (inter-cellular mobility). Although still fragmentary, our present view of the bacterial population dynamics derived from analysis of environmental samples supports the idea that MGEs are acting as a powerful evolutionary engine and regulator on such populations (2). It is therefore not surprising that all major large-scale sequencing projects now take into account the MGE contribution to the studied bacterial population, from the species level up to an entire environmental niche. Specialized databases and analysis systems dedicated to MGEs are becoming crucial for capturing and describing those highly diverse and fast evolving populations and for better understanding the mechanisms responsible for such diversity. However, few databases dedicated to prokaryotic MGEs have been developed so far (3–6) and the biological information related to MGEs remains sparse.

When the ACLAME database was first released in 2004 (3), the number of proteins/genes with a proper annotation was representing only a tiny fraction of the MGE sequence space. Knowing that there is not a single gene common to all phages or all plasmids, discriminating MGEs from bacteria in metagenomic data for instance remains a challenge. A set of structured and well-curated annotations appears as an important step in the set up of proper MGE analysis in large-scale sequencing projects.

The developments of the ACLAME database and associated analysis tools are intended to: (i) collect, in a dedicated system, all known MGE sequences, (ii) provide a platform for analyzing MGE diversity from a global scale down to specific groups of MGEs and (iii) provide tools for the detection of new MGEs integrated in bacterial genomes. This update article will describe all the new features added to the database and web site developed since the first publication in 2004.

*To whom correspondence should be addressed. Tel: +32 2 650 5499; Fax: +32 2 650 5425; Email: raphael@bigre.ulb.ac.be

UPDATED CONTENT

The last release of the ACLAME database (version 0.4) contains information on 457 bacteriophage genomes, 1109 plasmids and 760 prophages. With every ACLAME update, individual proteins encoded by the MGEs are used as a query sequence with BLASTP (7) (*E*-value threshold of 0.001) against the non-redundant (nr) database from NCBI (8), the SwissProt database (9) and the SCOP database (10) to collect a maximum number of similar sequences. These are used as support for the annotation process. Protein families are built from the MGE encoded proteins, using a slightly modified procedure compared to the original publication. The first step consists in performing pairwise comparisons between all the protein sequences. The BLASTP program was initially used and has been replaced by the SSEARCH program from the FASTA3 package (11) with an *E*-value threshold of 0.001. SSEARCH computes and screens the full similarity matrix during sequence comparison, which allows for a better discrimination between sequences sharing widespread domains and for a better overall alignment quality. The second step, unchanged, runs the clustering algorithm MCL (12) (inflation factor 1.2) with the list of paired sequences (each sequence defining a node in the graph) and using the *E*-values for defining the graph weighted edges. The ACLAME families are continuously manually annotated, now using only the GeneOntology (13) when it provides appropriate terms and our own ontology, called MeGO [http://aclame.ulb.ac.be/mego (14)] whenever MGE specific terms are required for annotating the protein families.

A new 'Prophages' category added in ACLAME version 0.4 represents a collection of high-quality predicted prophages retrieved from the Prophinder database. Prophinder (15) is a heuristic method, which relies on phage data in ACLAME, and is capable of detecting many prophages in completely sequenced bacterial genomes, while producing few false positives (sensitivity of 78% and positive predictive value of 95% from the published benchmark). Clustering of the combined set of phages and 1058 predicted prophages following a procedure described earlier (16) served to identify clusters composed exclusively of predicted prophages, which were further manually inspected. After this semiautomatic curation, we kept a set of 760 predictions, which make up the new 'Prophages' category. An additional category called 'Viruses and prophages' was built, combining phage and prophage data. This second new category turned out as a valuable source of information for improving phage and prophage annotations. In addition, the protein families obtained from the 'Viruses and prophages' category were used for further bioinformatics analysis (see below) allowing an unprecedented view of the intricate relationships among phages and prophages.

RETICULATE CLASSIFICATIONS

MGEs are well known for exchanging genes between them and with their hosts and hence their genomes are mosaics of groups of genes that remain more or less conserved within groups of related MGEs, while the rest can have multiple origins (17–20). Capturing lateral gene transfer events, the mosaicism and genome rearrangement events observed in MGE populations and inferring evolutionary links between MGEs requires novel approaches distinct from the traditional tree-like based methods (21). We proposed a graph-based method that was first used for the construction of a reticulate classification of phages (16). The procedure was applied on all ACLAME MGE categories and the resulting classifications are now accessible in the ACLAME database and web site. A major added value is the immediate access to the direct neighbors of any MGE in the graph which are the most similar entries of a given MGE, either as HTML tables or localized in the entire classification space, displayed as a graph using the interfaced Cytoscape application (22). The MGE view section on the ACLAME web site now provides the new option 'Classification'. On the 'Classification' page, users can display the entire graph or only the most closely related MGEs by using the Cytoscape button. On the same page, a matrix view summarizes proteins in common between the viewed MGE and its closest relatives with a full description of the common proteins per MGE further down in the page. Figure 1 gives an overview of the reticulate classification of phage Mu (mge:320).

EVOLUTIONARY COHESIVE MODULES

As mentioned earlier, no single gene is common to all plasmids or all (pro)phages, however, genetic modules are shared between related groups of MGEs. These genetic modules can be responsible for specific functions, hence their identification would be valuable for a comprehensive classification and typing of MGEs. Based on the idea that genes with similar profile occurrence across genomes are likely involved in the same function (23), we developed a methodology capable of defining evolutionary cohesive modules (ECMs) (16). ECMs were generated for all ACLAME categories and stored in the ACLAME database. They are accessible on the web site through the MGE viewer, following the link 'ECM'. Figure 2 shows an example of ECMs containing proteins belonging to phage Mu on the ACLAME web site. A detailed analysis of ECMs obtained from the 'Viruses and prophages' category will be published elsewhere.

ACCESS AND INTERFACE

The ACLAME database is now running under the PostgreSQL relational database management system (RDBMS) version 8.3 (http://www.postgresql.org). Satellite databases have been integrated for cross-referenced data. The GeneOntology and MeGO databases are installed under the MySQL RDBMS version 5.1 (http://www.mysql.com). They are used to retrieve functional terms displayed on the ACLAME web site. The Prophinder database is installed under the same PostgreSQL RDBMS and is used for accessing all the details related to the predicted prophages added in

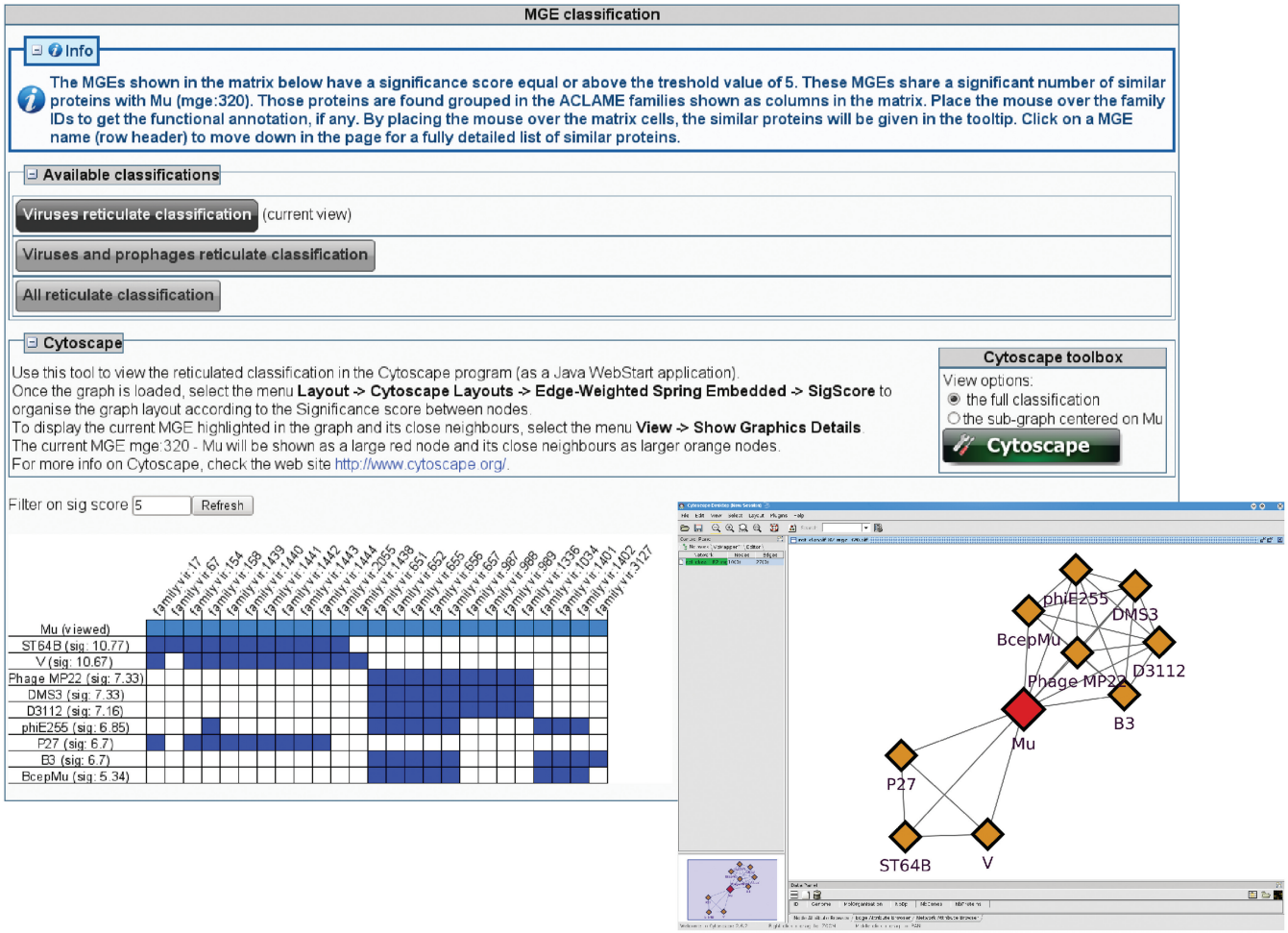

Figure 1. Classification section on the web site for transposable phage Mu. In the matrix view on the left, the columns represent the protein families where the MGE (Mu in this case) encoded proteins are found. The first row represents phage Mu and the following rows correspond to the most closely related MGEs. A filled cell (blue) corresponds to a protein of the MGE in the given row and that is found in the same family of the Mu protein present in the same column. The number in parenthesis next to the MGE name is the significance (sig) score defining the degree of similarity with Mu. The graph on the right corresponds to the output of the 'sub-graph centered on Mu' button in the Cytoscape toolbox. The graph shows the relationships between the seven mutator phages present in ACLAME version 0.4. Most shared proteins are involved in replication and head morphogenesis and structure (see below). Mu, a Myoviridae as BcepMu and PhiE255, is the only one in the group to share tail proteins with a set of non-mutator phages, namely *E. coli* phage P27, *Salmonella enterica* phage ST64B and *Shigella flexneri* phage V, all present in the matrix and graph views.

the ACLAME database. Additional databases under PostgreSQL contain the NCBI taxonomy and GenBank data (8), providing efficient access to the taxonomical and other genomic data cross-referenced with the MGEs and their hosts.

From the web site, most data available in the database are downloadable, either as simple tab-delimited text files or as Excel spreadsheets and all the sequences can be downloaded in FASTA formatted text files.

The web interface is continuously improved with the aim of facilitating the navigation. A simple text query form allows for searching most of the database content. The BLAST search interface has been extended and improved. It allows for a simple querying of the ACLAME sequences, returning a number of information such as for each hit sequence, the functional annotation, the MGE, host(s) and protein families it belongs to. A new view has been added to display all the hits grouped by MGE, especially useful to easily spot closest relatives when submitting all the proteins encoded by a newly sequenced MGE. A tool for downloading specific parts of the results in a tab-delimited text file format facilitates the import on a local computer.

FUTURE DIRECTIONS

The implementation of the reticulate classifications and the ECMs on the ACLAME web site was our major goal for version 0.4. Additional tools required for accessing and using the classifications and the modules on the web site are currently being developed. Functional annotation of the ECMs, a source of unique set of genes associated with particular functional classes, is under way. ECMs can be used to automatically identify

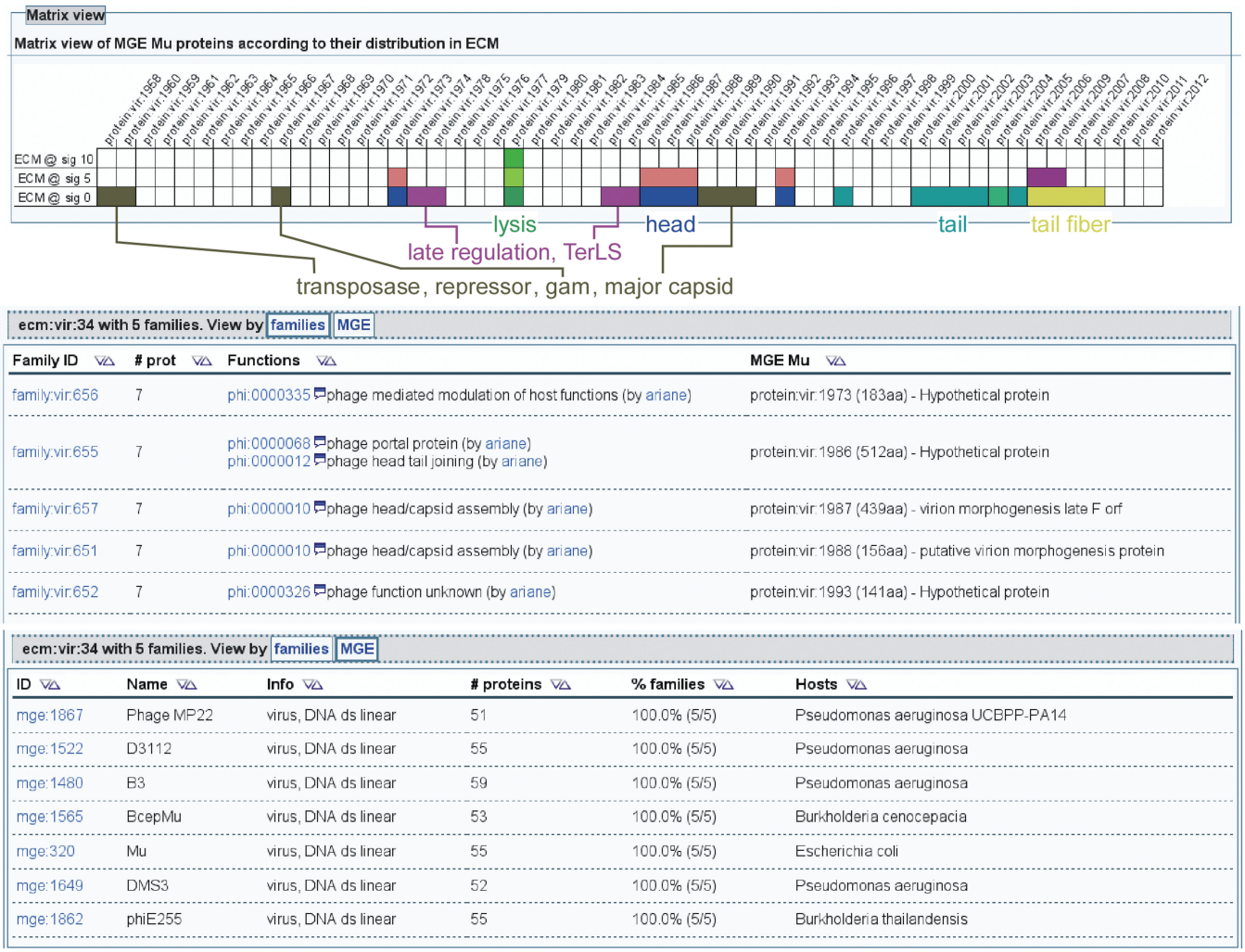

Figure 2. ECM section on the web site for transposable phage Mu. In the matrix view on the top of the figure, columns correspond to Mu encoded proteins as ordered in the phage genome. Rows show the ECMs obtained at three sig thresholds, 10, 5 and 0, with a different color code for each row since ECMs composition change according to the sig threshold. Proteins with similar colors belong to families that are part of the same ECM. The functions of the proteins in the ECMs have been added on this figure. On the website, each ECM composition, in term of protein families, can be viewed as tables with the MGE (Mu) proteins occurrence highlighted in a specific column (middle table on the figure). The table content can be changed to display the MGEs contribution, in term of proteins distribution, to the families present in the ECM (lower table).

particular 'life-style' or host specificity among (newly sequenced) MGEs. The reticulate classification is powerful for capturing evolutionary relationships among MGEs and although the aim is not to move towards a taxonomical classification, it should be valuable both to validate known taxonomic classes [for instance those in the ICTV taxonomy of viruses (24)] and to suggest new classes of MGEs. Combining the information in ECMs with the reticulate classification should allow for the prediction of functions from proteins with so far unknown function. Indeed, the co-occurrence of such proteins with others having curated functional annotations and belonging to a group of related MGEs could hint to their potential function and eventually make them ideal targets for experimental validation.

The open annotation platform on the ACLAME web site will provide individuals or groups with the possibility to register as official maintainer of specific MGEs and contribute to a continuous update of the associated information. This opens the way for redistributing in a human and computer readable format any valuable information so far stated by experimentalists only in publications or stored in a document on their personal computer. We see this option as one of the best ways to ensure high quality annotations, which is the very foundation of any bioinformatics studies.

ACKNOWLEDGEMENTS

The authors would like to acknowledge and give credit to the user community for contributing to the improvements of the database and web site content. They are grateful as well to J.C. Alonso, S. Casjens, G. Christie, E. Haggard-Ljungquist, G. Hatfull, J. Klumpp, I. Molineux, M. Salas and M. Yarmolinsky who provided valuable information for the annotations of reference phages.

FUNDING

European Space Agency (ESA-PRODEX) and the Belgian Science Policy (Belspo) through the MISSEX project (PRODEX agreements No. C90254), the Fonds de la Recherche Scientifique Médicale (FRSM) and the Belgian Program on Interuniversity Attraction Poles, initiated by the Belgian Federal Science Policy Office, project P6/25 (BioMaGNet). The BiGRe laboratory is a partner of the BioSapiens Network of excellence funded under the sixth Framework program of the European Communities (LSHG-CT-2003-503265). Funding for open access charge: Convention between the Belgian Nuclear Research Centre SCK/CEN and the Free University of Brussels (ULB), (contract number: FH1314000001).

Conflict of interest statement. None declared.

REFERENCES

1. Frost,L.S., Leplae,R., Summers,A.O. and Toussaint,A. (2005) Mobile genetic elements: the agents of open source evolution. *Nat. Rev. Microbiol.*, **3**, 722–732.
2. Rohwer,F. and Thurber,R.V. (2009) Viruses manipulate the marine environment. *Nature*, **459**, 207–212.
3. Leplae,R., Hebrant,A., Wodak,S.J. and Toussaint,A. (2004) ACLAME: a CLAssification of Mobile genetic Elements. *Nucleic Acids Res.*, **32**, D45–D49.
4. Siguier,P., Perochon,J., Lestrade,L., Mahillon,J. and Chandler,M. (2006) ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res.*, **34**, D32–D36.
5. Moura,A., Soares,M., Pereira,C., Leitão,N., Henriques,I. and Correia,A. (2009) INTEGRALL: a database and search engine for integrons, integrases and gene cassettes. *Bioinformatics*, **25**, 1096–1098.
6. Molbak,L., Tett,A., Ussery,D.W., Wall,K., Turner,S., Bailey,M. and Field,D. (2003) The plasmid genome database. *Microbiology*, **149**, 3043–3045.
7. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
8. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2009) GenBank. *Nucleic Acids Res.*, **37**, D26–D31.
9. The Universal Protein Resource (UniProt). (2009) *Nucleic Acids Res.*, **37**, D169–D174.
10. Andreeva,A., Howorth,D., Chandonia,J.M., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.*, **36**, D419–D425.
11. Pearson,W. and Lipman,D. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
12. Enright,A.J., Van Dongen,S. and Ouzounis,C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.
13. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
14. Toussaint,A., Lima-Mendez,G. and Leplae,R. (2007) PhiGO, a phage ontology associated with the ACLAME database. *Res. Microbiol.*, **158**, 567–571.
15. Lima-Mendez,G., Van Helden,J., Toussaint,A. and Leplae,R. (2008) Prophinder: a computational tool for prophage prediction in prokaryotic genomes. *Bioinformatics*, **24**, 863–865.
16. Lima-Mendez,G., Van Helden,J., Toussaint,A. and Leplae,R. (2008) Reticulate representation of evolutionary and functional relationships between phage genomes. *Mol. Biol. Evol.*, **25**, 762–777.
17. Lawrence,J.G. (2002) Gene transfer in bacteria: speciation without species? *Theor. Popul. Biol.*, **61**, 449–460.
18. Norman,A., Hansen,L.H. and Sørensen,S.J. (2009) Conjugative plasmids: vessels of the communal gene pool. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **364**, 2275–2289.
19. Toussaint,A. and Merlin,C. (2002) Mobile elements as a combination of functional modules. *Plasmid*, **47**, 26–35.
20. Lima-Mendez,G., Toussaint,A. and Leplae,R. (2007) Analysis of the phage sequence space: the benefit of structured information. *Virology*, **365**, 241–249.
21. Rohwer,F. and Edwards,R. (2002) The Phage Proteomic Tree: a genome-based taxonomy for phage. *J. Bacteriol.*, **184**, 4529–4535.
22. Killcoyne,S., Carter,G.W., Smith,J. and Boyle,J. (2009) Cytoscape: a community-based framework for network modeling. *Methods Mol. Biol.*, **563**, 219–239.
23. Pellegrini,M., Marcotte,E.M., Thompson,M.J., Eisenberg,D. and Yeates,T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.
24. Büchen-Osmond,C. (1997) Further progress in ICTVdB, a universal virus database. *Arch. Virol.*, **142**, 1734–1739.