INSTITUT PASTEUR

Research in Microbiology 158 (2007) 567–571

www.elsevier.com/locate/resmic

PhiGO, a phage ontology associated with the ACLAME database

Ariane Toussaint*, Gipsi Lima-Mendez, Raphaël Leplae

Service de Conformation de Macromolécules Biologiques et de Bioinformatique, Université Libre de Bruxelles, ULB, Bvd du Triomphe, Campus Plaine – CP 263, B1050 Bruxelles, Belgium

Received 19 March 2007; accepted 10 May 2007
Available online 21 May 2007

Abstract

As is the case for other genomes and metagenomes, complete nucleotide sequences of bacteriophages and archaeviruses have become increasingly numerous and require robust annotation tools. We present here the first version of a phage ontology, PhiGO, which should contribute to more informational annotation of gene products in phage and prophage sequences. PhiGO uses the Gene Ontology schema, the de facto standard for describing knowledge about gene products across many databases.

Keywords: Bacteriophage; Genomics; Ontology

1. Introduction

Biology has entered an era in which the amount of data produced by a single experiment is no longer manually manageable. Computational analysis has become unavoidable, resulting in the need for new standards in the formatting of data so that they can be easily distributed on the Internet and accessed by computer programs. Standard formats have therefore been developed for many types of high throughput data, such as nucleotide and amino acid sequences and their annotations. Sequences are deposited in a GenBank/EMBL format, which includes various fields such as those describing gene coordinates, coding sequences and gene products, just to mention a few. Such descriptions are not sufficient to capture the biology of the sequence, the gene, its product or their context, i.e. the biological entity to which they belong. This has promoted the development of ontologies such as MultiFun [11], developed for the annotation of bacterial genome sequences, and Gene Ontology (GO, http://www.geneontology.org/, [5]), originally developed in the context of the *Drosophila* genome project. Since then, GO has become the de facto standard for describing the principal attributes (molecular function, biological processes and cellular components) of gene products across numerous databases. It succeeds in the major aim of an ontology by providing a common, shared understanding of the concepts used to describe those attributes (with, at present, a major focus on some eukaryotic organisms). It does this by providing terms used to label those concepts as well as natural language definitions of those terms. The more recent SO ontology [2] builds up a description of the parts of a genomic annotation.

Bacteriophage and archaevirus genomics is not particularly prominent either in terms of the number of sequences available or data structure and description. The phage section of the ACLAME database for storage and annotation of prokaryotic mobile genetic elements [7] is attempting to fill in some of these gaps. ACLAME provides access to a set of phage and plasmid protein families, based on sequence similarity, which have been connected to a provisional functional annotation based on a list of functions built for that purpose (http://aclame.ulb.ac.be/functions). The phage-related functions in that list have now been amended and updated into the PhiGO ontology, using the GO schema.

* Corresponding author. Tel.: +32 2 650 5499; fax: +32 2 650 5425.
E-mail address: ariane@scmbb.ulb.ac.be (A. Toussaint).

0923-2508/$ - see front matter © 2007 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.resmic.2007.05.002

2. Materials and methods

Version 0.2 of the ACLAME database, including procedures used to build up protein families and the list of functions, has been described [7–9]. Phage terms and definitions have been assembled in the OBO file format [4,5] using the Java program OBO-Edit version 1.002 ([5], accessible from http://www.geneontology.org/). The OBO file has been loaded into the MySQL relational database version 1.1 using the GO-DEV package (http://amigo.geneontology.org/dev/). The AmiGO web interface has been installed on the ACLAME web server and is accessible at http://aclame.ulb.ac.be/Classification/phage_functions.html.

3. Results and discussion

Different clustering procedures have been implemented to group proteins annotated on sequenced phage genomes into families [6,9,10], where families are defined as an ensemble of similar sequences sharing a common function. In general, around 50% of the sequences cluster into families of three or more members. The ACLAME phage protein families of three or more members were annotated with functions. These functions provided the base for a list of terms in the PhiGO ontology, defining a phage life cycle formalized as a directed acyclic graph (DAG [4]). A ''directed graph'' is a structure with ''nodes'' and ''edges'', the latter being ordered pairs of nodes. Nodes and edges are ''labeled'', the first with the term denoting the class they stand for and the second with the kind of relationship that relates corresponding classes (e.g. ''is-a'' or ''part-of''). The graph (Fig. 1) is acyclic because paths via edges are only going ''forward''. The term labeling a node refers to this node and all of its children.

3.1. Phage ontology terms

A set of definitions for terms related to phage components, lytic and lysogenic cycles (biological processes here called ''phage processes'') and molecular functions (here called ''phage activities'') have been assembled under the OBO format. Users can navigate in the ontology using the AmiGO web interface, and the OBO file can be downloaded from the ACLAME website (http://aclame/perl/Services/amigo_phage/go.cgi). Examples of the views are given in Fig. 2.

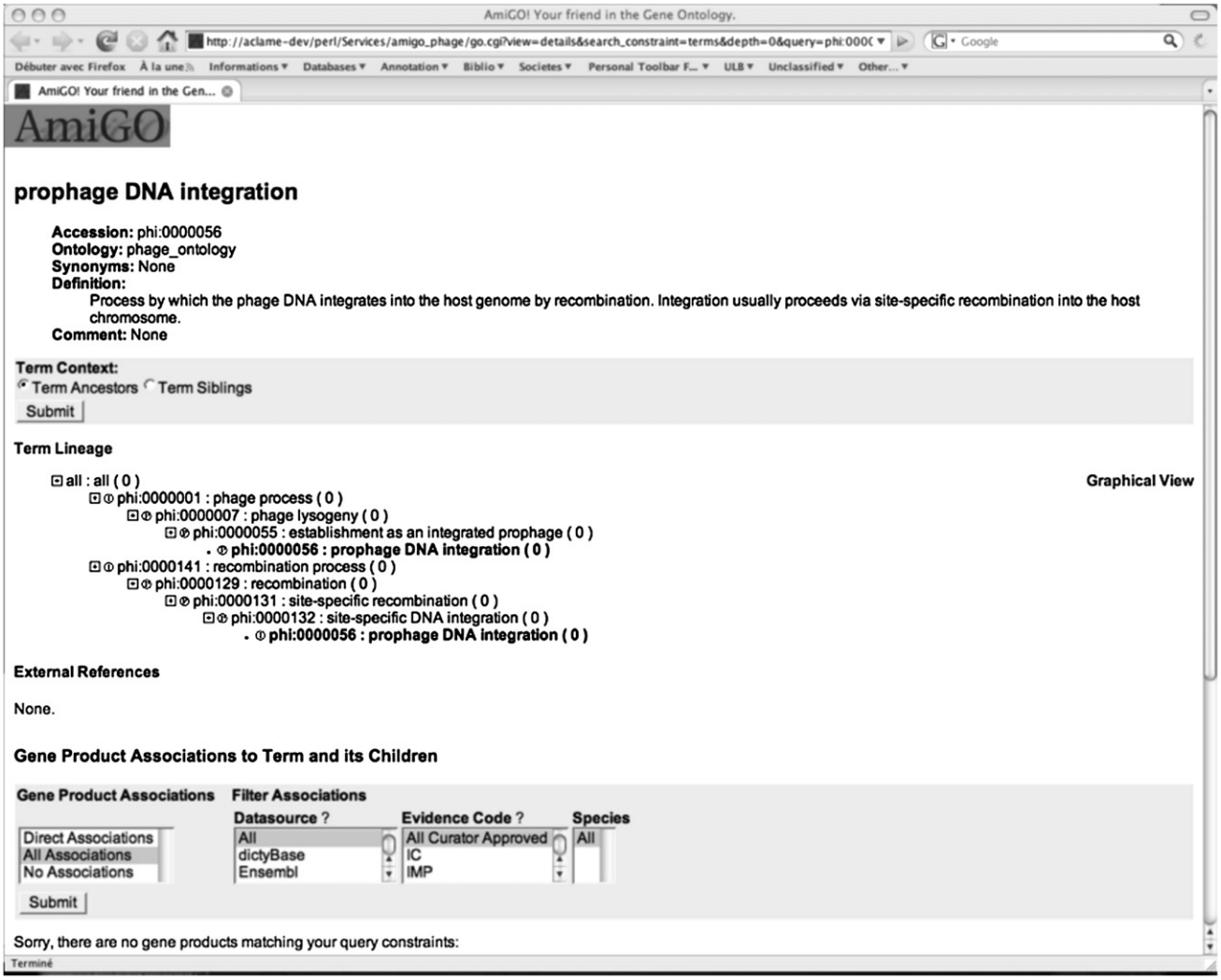

Fig. 1. View of the term ''prophage DNA integration'' in the AmiGO format. ''p'' and ''i'', respectively stand for ''part-of'' and ''is-a'' relationships.

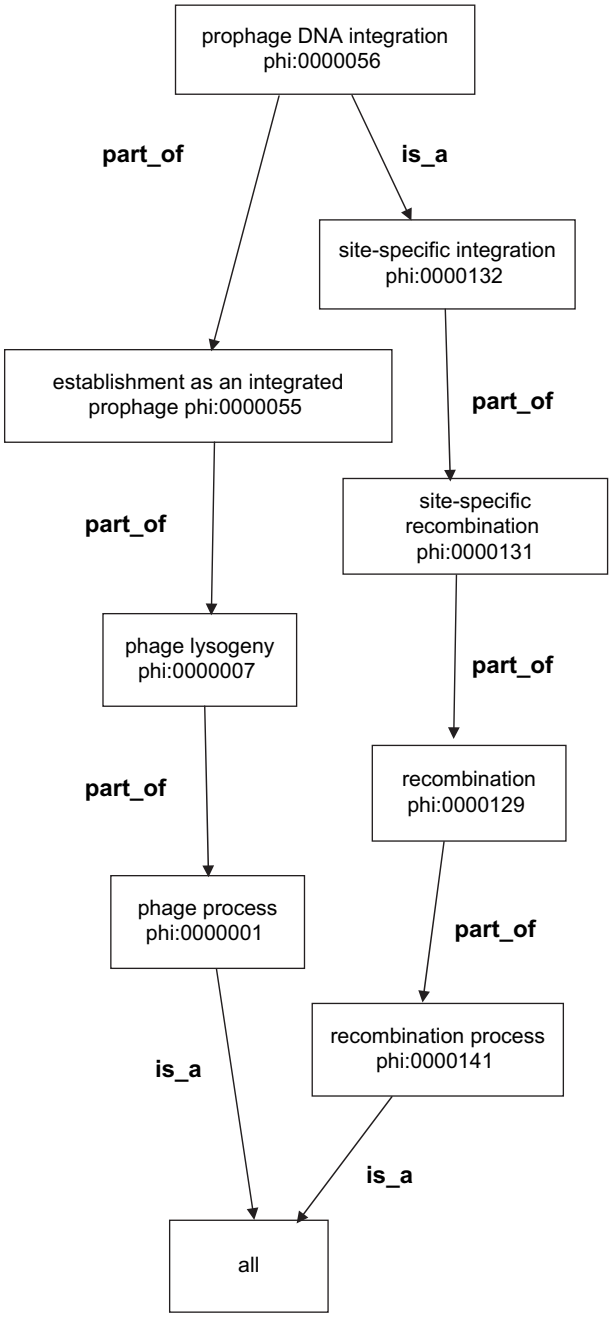

Fig. 2. DAG (directed acyclic graph) representation of the term ‘‘prophage DNA integration’’. Nodes in the graph are ‘‘terms’’ while edges are relationships (part-of or is-a).

The present list of terms in phage ontology covers rather generic phage features and it should be further developed to capture all types of phages and their specificities, keeping in mind the requirement for a common vocabulary which can nevertheless include a series of synonyms for any term.

3.2. ACLAME phage ontology as an aid to phage sequence annotation

ACLAME provides a classical ‘BLASTP’ search facility, which takes a set of query protein sequences and retrieves similar protein sequences from the ACLAME database. For each protein in ACLAME defined as similar, the output provides the usual sequence alignment, a link to the family containing that protein and the ACLAME function(s) assigned to that family (Fig. 3A).

Combined with the Prophinder tool for the prediction of prophages in bacterial genome sequences (http://aclame.ulb.ac.be/prophinder), phage ontology should facilitate automatic annotation of prophages and their gene products in bacterial genome sequences. At present, prophages predicted by Prophinder, Phage-Finder (another prophage prediction

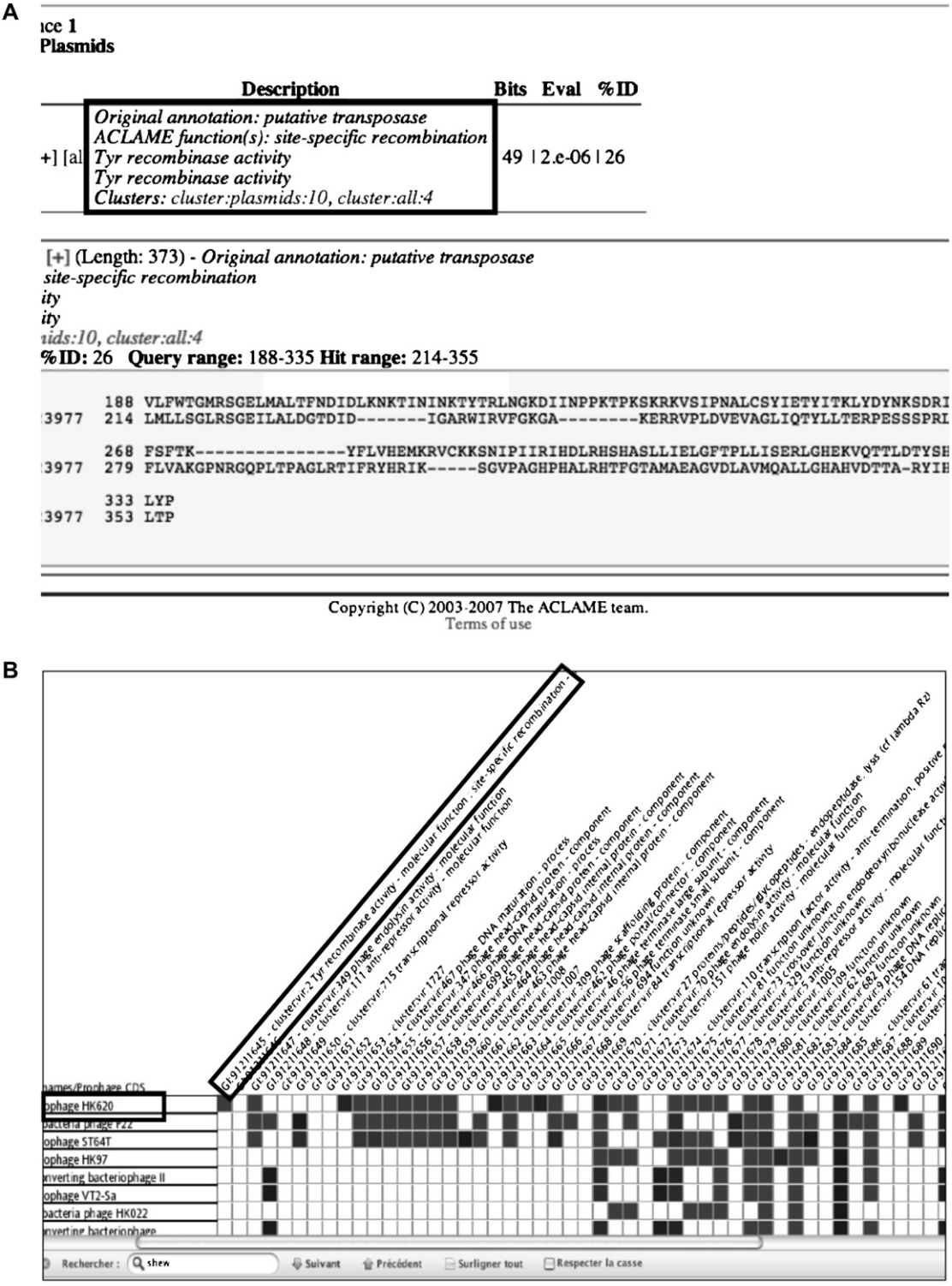

Fig. 3. (A) ACLAME BLASTP output on a tyrosine recombinase protein. The proposed annotation (in the black rectangle) appears as that appended to the most related protein in the database, here protein:vir3724. (B) Heat map (partial view) of a prophage predicted with Prophinder in the *Escherichia coli* UTI89 genome sequence. Each CDS from the predicted prophage is displayed associated with the annotation of the most related protein in ACLAME. The first column shows the name of the phage whose genome contains that protein.

tool, [3]) and manually [1] already display, for each CDS of the bacterial genome sequence that has a hit in ACLAME, the function(s) linked to the closest relative protein in the database (Fig. 3B).

Acknowledgments

We wish to thank M. Ashburner and J. Lomax for introducing us to GO and OBO-Edit, S. Casjens and I. Molineux for

their contribution to the assembly of phage terms and definitions, and G. Chaconas, A. Landy and G. Hatfull for suggesting corrections and new definitions. This work was supported by ESA-PRODEX (contract C90254), the Belgian Fonds de la Recherche Fondamentale Collective (FRFC) and the Université Libre de Bruxelles (ULB). GLM is a fellow from the Fonds Xenophilia, ULB. Any suggestions, comments, corrections and additional terms can be sent to ariane@scmbb.ulb.ac.be, with the PhiGO term in the subject line, or through the ACLAME forums.

References

[1] Casjens, S. (2003) Prophages and bacterial genomics: what have we learned so far? Mol. Microbiol. 49, 277–300.

[2] Eilbeck, K., Lewis, S.E., Mungall, C.J., Yandell, M., Stein, L., Durbin, R., Ashburner, M. (2005) The sequence ontology: a tool for the unification of genome annotations. Genome Biol. 6, R44.

[3] Fouts, D.E. (2006) Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. Nucleic Acids Res. 34, 5839–5851.

[4] Gene Ontology Consortium (2001) Creating the Gene Ontology resource: design and implementation. Genome Res. 11, 1425–1433.

[5] Harris, M.A., Clark, J., Ireland, A., Lomax, J.V., Ashburner, M.V., Foulger, R.V., Eilbeck, K.V., Lewis, S.V., Marshall, B., Mungall, C., Richter, J., Rubin, G.M., Blake, J.A., Bult, C., Dolan, M., Drabkin, H., Eppig, J.T., Hill, D.P., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J.M., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S., Fisk, D.G., Hirschman, J.E., Hong, E.L., Nash, R.S., Sethuraman, A., Theesfeld, C.L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S.Y., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee, V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E.M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., de la Cruz, N., Tonellato, P., Jaiswal, P., Seigfried, T., White, R.Gene Ontology Consortium (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 32, D258–D261.

[6] Hatfull, G.F., Pedulla, M.L., Jacobs-Sera, D., Cichon, P.M., Foley, A., Ford, M.E., Gonda, R.M., Houtz, J.M., Hryckowian, A.J., Kelchner, V.A., Namburi, S., Pajcini, K.V., Popovich, M.G., Schleicher, D.T., Simanek, B.Z., Smith, A.L., Zdanowicz, G.M., Kumar, V., Peebles, C.L., Jacobs Jr., W.R., Lawrence, J.G., Hendrix, R.W. (2006) Exploring the mycobacteriophage metaproteome: phage genomics as an educational platform. PLoS Genet. 2(6), e92.

[7] Leplae, R., Hebrant, A., Wodak, S.J., Toussaint, A. (2004) ACLAME: a CLAssification of Mobile genetic Elements. Nucleic Acids Res. 32, D45–D49.

[8] Leplae, R., Lima-Mendez, G., Toussaint, A. (2006) A first global analysis of plasmid encoded proteins in the ACLAME database. FEMS Rev. 30, 980–994.

[9] Lima-Mendez G., Toussaint, A. and Leplae, R. Analysis of the phage sequence space: the benefit of structured information. Virology, in press.

[10] Liu, J., Glazko, G., Mushegian, A. (2006) Protein repertoire of double-stranded DNA bacteriophages. Virus Res. 117, 68–80.

[11] Serres, M.H., Riley, M. (2000) MultiFun, a multifunctional classification scheme for *Escherichia coli* K-12 gene products. Microb. Comp. Genomics 5, 205–222.