The Gypsy Database (GyDB) of mobile genetic elements: release 2.0

Carlos Llorens1,*, Ricardo Futami1, Laura Covelli1, Laura Domínguez-Escribá1, Jose M. Viu1, Daniel Tamarit2, Jose Aguilar-Rodrı´guez2, Miguel Vicente-Ripolles3, Gonzalo Fuster1, Guillermo P. Bernet1,4, Florian Maumus5, Alfonso Munoz-Pomer1,3, Jose M. Sempere3, Amparo Latorre2,6 and Andres Moya2,6

1Biotechvana, Parc Científic, Universitat de València, Calle Catedrático José Beltrán 2, 46980 Paterna (València), 2Unidad Mixta de Investigación en Genómica y Salud del Centro Superior de Investigación en Salud Pública (CSISP)-Universitat de València (Instituto Cavanilles de Biodiversidad y Biología Evolutiva), Avenida de Cataluña 21, 46020 València, 3Departamento de Sistemas Informáticos y Computación (DSIC), Universitat Politècnica de València, Camino de Vera S/N, 46022 València, 4Instituto Valenciano de Investigaciones Agrarias (IVIA), Carretera Moncada-Naquer, Km 4.5, 46113, Moncada (València), Spain, 5Institut Jean-Pierre Bourgin, INRA Centre de Versailles-Grignon, Route de Saint-Cyr, 78026 Versailles, France and 6CIBER en Epidemiología y Salud Pública (CIBEResp), Parc de Recerca Biomèdica de Barcelona, Calle Doctor Aiguader 88 1ª Planta, 8003 Barcelona, Spain

Received September 2, 2010; Revised October 11, 2010; Accepted October 13, 2010

ABSTRACT

This article introduces the second release of the Gypsy Database of Mobile Genetic Elements (GyDB 2.0): a research project devoted to the evolutionary dynamics of viruses and transposable elements based on their phylogenetic classification (per lineage and protein domain). The Gypsy Database (GyDB) is a long-term project that is continuously progressing, and that owing to the high molecular diversity of mobile elements requires to be completed in several stages. GyDB 2.0 has been powered with a wiki to allow other researchers participate in the project. The current database stage and scope are long terminal repeats (LTR) retroelements and relatives. GyDB 2.0 is an update based on the analysis of Ty3/Gypsy, Retroviridae, Ty1/Copia and Bel/Pao LTR retroelements and the Caulimoviridae pararetroviruses of plants. Among other features, in terms of the aforementioned topics, this update adds: (i) a variety of descriptions and reviews distributed in multiple web pages; (ii) protein-based phylogenies, where phylogenetic levels are assigned to distinct classified elements; (iii) a collection of multiple alignments, lineage-specific hidden Markov models and consensus sequences, called GyDB collection; (iv) updated RefSeq databases and BLAST and HMM servers to facilitate sequence characterization of new LTR retroelement and caulimovirus queries; and (v) a bibliographic server. GyDB 2.0 is available at http://gydb.org.

INTRODUCTION

Mobile genetic elements (MGEs) are ubiquitous, autonomous genetic units that often constitute a significant part of their host genomes. It is commonly accepted that mobile DNA elements are powerful vectors for disease and evolution, from which distinct host genes have evolved during the history of life (1,2). The emergence and subsequent role played by viruses and MGEs in the history of life is an exciting topic that requires further investigation. In this respect, researchers aim to discern relevant aspects of the molecular changes responsible for various characteristics in organisms related to horizontal transfer, infection and disease. Among the distinct initiatives launched with the aim of investigating the diversity of MGEs (see for example 3–5) was the Gypsy Database (GyDB) of MGEs (6), a research project devoted to the evolutionary dynamics of viruses and MGEs (and their related host proteins), which was launched in 2008. The GyDB project is a highly
informative database established within an evolutionary context of classification, where one piece of research delivers one conclusion that drives individuals towards another goal. The most captivating aspect of this project is that a share of our efforts are dedicated to the interpretation of analyses, paying particular attention to non-redundant elements displaying a certain degree of distance and investigating how they can be collectively aligned or related, in terms of protein domain architecture, with other lineages and elements. Because of the impressive molecular diversity of viruses and MGEs, the GyDB is a long-term project that has been arranged in a database in continuous progression, and must be achieved in stages. The current database stage and scope is retroviruses and retrotransposons with long terminal repeats (LTR retroelements) and their relatives. Following the outline of the earlier release (the study of Ty3/Gypsy and Retroviridae LTR retroelements), this article presents the GyDB update based on the phylogenetic evaluation of the most representative LTR retroelement families and the plant caulimoviruses. This update, called GyDB 2.0, is available at http://gydb.org and includes sequence phylogenetic classification in addition to significant bioinformatic improvements. In particular, the new infrastructure implements a wiki management system constructed with the aim of promoting a world-wide community of researchers collaborating in the analysis and classification of MGEs and viruses inhabiting (or circulating in) living organisms.

THE UPDATE: NEW FEATURES

GyDB 2.0 consists of 1234 web pages addressing the phylogenetic study of Ty3/Gypsy, Retroviridae, Ty1/Copia and Bel/Pao LTR retroelement. Caulimoviruses (Caulimoviridae) are formally plant DNA pararetroviruses, but they were considered in GyDB 2.0 owing to their relationship with LTR retroelements based on the common gag/coat and pol regions [for more details, see (7) and references therein]. Table 1 summarizes the topics addressed in this update, as well as the servers and database sections it offers. The sequences on which GyDB 2.0 is based were retrieved from GenBank (8) and the methodologies employed were the same as those described earlier in references (6,7,9). At GyDB we evaluate the phylogenetic signal of classified distinct elements and create hidden Markov model (HMMs) profiles (10) per lineage and protein domain. In addition, the project is concerned with the evolutionary relationships between MGEs and their host genomes, based on the analysis of common protein families. In this regard, GyDB 2.0 focuses on two protein superfamilies including protein products commonly encoded by LTR retroelements and their host genomes; the chromodomain superfamily (11) and clan AA of aspartic peptidases (12,13). This second release is accompanied by bibliographic data-mining from PubMed databases hosted at the National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov).

Table 1. GyDB 2.0 new features: topics and contents

<table>
<thead>
<tr>
<th>Systems</th>
<th>Families</th>
<th>Lineages</th>
<th>Elements</th>
<th>Protein domains</th>
<th>Accessory proteins</th>
<th>LTRs</th>
</tr>
</thead>
<tbody>
<tr>
<td>LTR retroelements</td>
<td>Ty3/Gypsy</td>
<td>34</td>
<td>96</td>
<td>8</td>
<td>1</td>
<td>Yes</td>
</tr>
<tr>
<td>LTR retroelements</td>
<td>Ty1/Copia</td>
<td>19</td>
<td>69</td>
<td>8</td>
<td>–</td>
<td>Yes</td>
</tr>
<tr>
<td>LTR retroelements</td>
<td>Retroviridae</td>
<td>8</td>
<td>50</td>
<td>8</td>
<td>41</td>
<td>Yes</td>
</tr>
<tr>
<td>LTR retroelements</td>
<td>Bel/Pao</td>
<td>5</td>
<td>23</td>
<td>7</td>
<td>–</td>
<td>Yes</td>
</tr>
<tr>
<td>LTR retroelements</td>
<td>Caulimoviridae</td>
<td>6</td>
<td>30</td>
<td>10</td>
<td>27</td>
<td>No</td>
</tr>
<tr>
<td>Related families</td>
<td>Clan AA</td>
<td>35</td>
<td>323</td>
<td>1</td>
<td>–</td>
<td>No</td>
</tr>
<tr>
<td>Related families</td>
<td>Chromodomains</td>
<td>2</td>
<td>123</td>
<td>1</td>
<td>–</td>
<td>No</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Topics</th>
<th>Sections</th>
<th>Availability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Systematics</td>
<td>9</td>
<td>Side menu</td>
</tr>
<tr>
<td>Domains</td>
<td>14</td>
<td>Side menu</td>
</tr>
<tr>
<td>Database</td>
<td>8</td>
<td>Side menu</td>
</tr>
<tr>
<td>Servers</td>
<td>3</td>
<td>Top menu: BLAST, HMM, Literature</td>
</tr>
<tr>
<td>Wiki tools and utilities</td>
<td>3</td>
<td>Top menu</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Databases</th>
<th>Items</th>
<th>Sections</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genomes (full-length genomes)</td>
<td>271</td>
<td>sequences, BLAST search and RefSeq DBs</td>
</tr>
<tr>
<td>LTRs (nucleotide sequences)</td>
<td>413</td>
<td>sequences, BLAST search and RefSeq DBs</td>
</tr>
<tr>
<td>Cores (protein cores sequences)</td>
<td>1895</td>
<td>sequences, BLAST search and RefSeq DBs</td>
</tr>
<tr>
<td>HMMs</td>
<td>314</td>
<td>HMM profiles, GyDB collection</td>
</tr>
<tr>
<td>Multiple alignments</td>
<td>131</td>
<td>alignments, GyDB collection</td>
</tr>
<tr>
<td>Consensus sequences</td>
<td>314</td>
<td>MRC sequences, GyDB collection</td>
</tr>
<tr>
<td>Phylogenetic trees</td>
<td>70</td>
<td>trees, Phylogenies</td>
</tr>
<tr>
<td>Clan AA ancestral reconstruction</td>
<td>70</td>
<td>alignments, CAARD database</td>
</tr>
<tr>
<td>Literature</td>
<td>100797</td>
<td>references, Literature server</td>
</tr>
</tbody>
</table>

*We included caulimoviruses in the second release in view of their relationship with LTR retroelements based on the common gag/coat and pol region.*
nih.gov/) to document up to date information regarding the distinct classified elements.

DATABASE ORGANIZATION

GyDB 2.0 is deployed over a Linux-MySQL-Apache-PHP (LAMP) stack, with additional Ajax programming to minimize server responses to client browsers. The design is similar to that of the previous release but implements various changes on the web interface. As shown in Figure 1, the database organization is founded upon two major menus—a top menu and a side menu. The top menu allows access to the three servers:

(i) BLAST server; implements a BLAST search powered by the NCBI BLAST package (14), allowing protein and DNA comparisons with the GENOMES, LTRs and CORES databases. These databases collect the full-length genomes, the LTR sequences and all the protein sequences on which the second release is based, respectively.

(ii) HMM server; implements HMMER3 package (http://hmmer.janelia.org) and allows protein...
comparisons against a database of protein domain lineage-specific HMM profiles created based on the update. This server provides additional comparisons between HMM profiles and the aforementioned CORES database.

(iii) LITERATURE server; allows users to search bibliography of interest in the topic.

An additional new tool in GyDB 2.0 is its wiki, powered by the MediaWiki content management system (http://www.mediawiki.org/). This tool has been implemented to allow other users participate in the project by editing or creating topics. Accession to this wiki is free but it requires a subscription (registration). The rationale behind this choice is that edits are registered by date and author in order to credit contributions, and secondly, we have programmed a revision mechanism to review all changes constructively before making them public. The top menu includes three sections to log in and manage the distinct wiki resources. Finally, to the right of the top menu, GyDB 2.0 includes a text field to search the whole project under two modes (detailed in Figure 1).

The side menu divides the distinct GyDB sections into three major demarcations (emphasized with boxes in Figure 1). The first collects sections associated with the systematics applied at GyDB. The second implements information concerning the domains typically observed in the genomic structure of the elements we classify. The third demarcation offers free access to distinct databases, which are organized into three sections:

(i) Trees and Networks; consists of the collection of inferred phylogenetic trees based on distinct protein domains encoded by the classified elements, or based on their concatenation (when they are parts of polyproteins). Remarkably, inferred pol polyprotein phylogenies based on the concatenation of the protease, reverse transcriptase, RNaseH and integrase domains, are the major criterion for assigning phylogenetic levels at GyDB 2.0 [results introduced in (7)]. Phylogenetic trees provide links to the corresponding element page at GyDB 2.0. By clicking any element name in any tree an entry assigned to this element is opened. These tree image maps were created using Phylograph 1.0 (15). This section includes the clan AA reference database (CAARD) of ancestral maximum likelihood (ML) reconstructions (13) that has been implemented and maintained at GyDB.

(ii) GyDB collection (16) or the repository of multiple alignments, HMMs, and majority rule consensus (MRC) sequences offered at GyDB 2.0. When a deposited alignment, profile or MRC sequence is associated with a journal publication, its entry in the collection includes citation information.

(iii) REF SEQ DATABASES or the repository for downloading the databases (GENOMES, CORES and LTRs) implemented in the BLAST server.

Finally, a variety of links to other database initiatives relevant to the topic are included in the side menu.

FUTURE PERSPECTIVES

Sequencing projects constantly deliver new types of MGEs [for example (17–22)]; hence the classification of non-redundant elements based on their phylogenetic signal is an open issue at GyDB, and results in the preparation of new sections. For example, we are committed to improving the understanding of the diversity and evolutionary dynamics of MGEs in eukaryotic and prokaryotic organisms. In this regard of eukaryotic LTR retroelements (the current database scope), the sequence repertoire at GyDB with representative elements retrieved from recently sequenced marine secondary endosymbionts including the brown alga Ectocarpus siliculosus (heterokont) and the coccolithophore Emiliania huxleyi (haptophyte) will be implemented. In terms of other research topics in preparation, one concerns the construction of a server devoted to the study of the complete set of MGEs and repeats (the mobilome) of biological genomes. This server will be introduced with two forthcoming publications focusing on the LTR retroelements and their related transposases of the pea aphid Acyrthosiphon pisum genome [see (23)]. At the technical level, we are exploring the application of formal grammars and machine learning algorithms to automate, as far as possible, the management and classification of the sequence data. We are also committed to developing solutions for other non-trivial difficulties that arise with the growing size of the databases. Viruses and MGEs usually show different rates of evolution and high variability depending on the evaluated protein or region. Therefore, we aim to implement more than one method of phylogenetic reconstruction to offer the user different perspectives based on different methods (or the opportunity to upload updated phylogenies via the wiki). On the other hand, the traditional view of the origin and evolution of biological systems is that they are usually monophyletic, but such an assumption has been challenged by increasing evidence suggesting that natural evolution can frequently proceed by gradual and vertical means, in addition to distinct modular, saltatory and reticulate events (24–36).

In this respect, we are investigating appropriate protocols to combine phylogenetic inference with new tendencies in network biology [see also (7)].

ACKNOWLEDGEMENTS

We thank all the colleagues detailed in the list available at (http://gydb.org/index.php/Acknowledgments) for their support in contributing images of biological host organisms. We are also grateful to Senior NAR Editor Dr Michael Galperin and to the two anonymous reviewers for their constructive comments in improving this article. Finally we also thank Denys Wheatley and Angela Panther from Biomedes for copyediting of this article.

FUNDING

Centro de Desarrollo Tecnológico Industrial (CDTI) (grant IDI-20100007, partial); Empresa Nacional de Innovación, S.A (ENISA) (17092008, partial); IMPIVA
REFERENCES


