D448–D454 *Nucleic Acids Research, 2009, Vol. 37, Database issue* *Published online 9 October 2008*
doi:10.1093/nar/gkn684

ATGC: a database of orthologous genes from closely related prokaryotic genomes and a research platform for microevolution of prokaryotes

Pavel S. Novichkov[1,*], Igor Ratnere[2], Yuri I. Wolf[3], Eugene V. Koonin[3] and Inna Dubchak[1,2]

[1]Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, [2]Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598 and [3]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38A, 8600 Rockville Pike, Bethesda, MD 20894, USA

Received August 18, 2008; Revised September 22, 2008; Accepted September 23, 2008

ABSTRACT

The database of Alignable Tight Genomic Clusters (ATGCs) consists of closely related genomes of archaea and bacteria, and is a resource for research into prokaryotic microevolution. Construction of a data set with appropriate characteristics is a major hurdle for this type of studies. With the current rate of genome sequencing, it is difficult to follow the progress of the field and to determine which of the available genome sets meet the requirements of a given research project, in particular, with respect to the minimum and maximum levels of similarity between the included genomes. Additionally, extraction of specific content, such as genomic alignments or families of orthologs, from a selected set of genomes is a complicated and time-consuming process. The database addresses these problems by providing an intuitive and efficient web interface to browse precomputed ATGCs, select appropriate ones and access ATGC-derived data such as multiple alignments of orthologous proteins, matrices of pairwise intergenomic distances based on genome-wide analysis of synonymous and nonsynonymous substitution rates and others. The ATGC database will be regularly updated following new releases of the NCBI RefSeq. The database is hosted by the Genomics Division at Lawrence Berkeley National laboratory and is publicly available at http://atgc.lbl.gov

INTRODUCTION

The number of completely sequenced prokaryotic genomes is growing exponentially, with a doubling time of ~21 months (1). As of August 2008, 847 bacterial and 97 archaeal genomes have been published, and about 1900 genome projects are ongoing according to the GOLD database (2). The coverage of the prokaryotic world in sequence databases is growing both in breadth (new phyla and families) and in depth (many specific branches of archaea and bacteria are being sampled extensively). The breadth of coverage is critical to enhance our understanding of the diversity of bacteria and archaea, and for attempts to decipher deep evolutionary relationships. Conversely, depth of coverage of tight groups of microbes allows researchers to focus on microevolutionary mechanisms and clade-specific evolutionary processes. In particular, the analysis of sets of closely related microbial genomes provides for the possibility to assess the constancy of evolution rate (3) and selective pressure (4), determine the rate of adaptive evolution (5), identify horizontally transferred genes (6) and reconstruct the history of genome rearrangements (7,8).

A crucial condition for the success of these microevolutionary studies is the availability of an appropriate set of genomes separated by an evolutionary distance that is 'right' for the given task. Collecting a data set with the appropriate similarity parameters and other characteristics turns out to be a major hurdle in most of such studies. Given the current rate of genome sequencing, it is difficult to follow the progress of the field and to know *a priori* which of the available genome subsets meet the requirements. Furthermore, extraction of the required content (genomic alignments, families of orthologs, conserved strings of adjacent genes, etc.) from a selected set of genomes is a complicated and time-consuming process. Most comparative-genomic studies, implicitly or explicitly, rely upon comparisons within and across sets of genes that are considered to be orthologs, that is, genes derived from the same ancestral gene in the last common ancestor of the compared genomes (9). Reliable and

*To whom correspondence should be addressed. Email: psnovichkov@lbl.gov

© 2008 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from nar.oxfordjournals.org at Lawrence Berkeley Lab on April 21, 2011

complete identification of orthologs is a crucial condition of both the prediction of gene functions in newly sequenced genomes and elucidation of trends of microbial evolution. Despite more than a decade of progress (10–15), construction of the complete set of orthologs for a group of organisms remains a substantial technical problem, especially, when scaled up to a large collection of genomes.

Thus, resources containing precomputed, aligned sets of orthologous genes from groups of genomes separated by different distances could provide crucial aid to microbial comparative genomics and studies on microbial evolution. There are several community resources that allow access to all publicly available microbial genomes, their annotation and various types of analysis including the analysis of orthologous genes. Among these are the NCBI Genomes database (16), MicrobesOnline (17) and IMG (18); IMG, in particular, provides DNA VISTA alignments (19) for several manually selected groups of closely related genomes. However, to our knowledge, none of the major repositories of genomic information and analytical tools has the capability to comprehensively analyze closely related genomes as a group.

Here we describe a database of microbial orthologous gene sets optimized for microevolutionary research. These sets are based on groups of closely related species and strains of bacteria and archaea that were dubbed Alignable Tight Genomic Clusters (ATGC). ATGC-derived sets of orthologs rely on extensive synteny between the genomes to boost the reliability of ortholog identification. Among other functionalities, the ATGC database provides access to pairwise whole-genome alignments of the clusters of microbes where genomes are selected based on a custom set of parameters. The database interface facilitates the selection of subsets of the data customized for the specific requirements of particular research projects.

DATABASE CONSTRUCTION AND STRUCTURE

Data source

The current version of the ATGC website is based on RefSeq release 26 (4 November 2007). The microbial genomes were downloaded from NCBI ftp site ftp://ftp.ncbi.nlm.nih.gov/refseq/release/microbial. RefSeq sequences of plasmids and transposons were excluded from the analysis, so only the chromosome sequences were used to build the database. Both completely sequenced, finished genomes and genomes that are not yet labeled as complete are included considering that since, in many cases, incomplete genomes could be also informative. Information on the status of each genome is provided, so a user can decide whether to use incomplete genomes for a particular project.

Database content and organization

The primary object of the database is a cluster of closely related genomes, which we denote Alignable Tight Genomic Cluster (ATGC). Genomes are considered to be closely related and are included in an ATGC if they satisfy two criteria. First, genomes are required to form a tight cluster, which means that they should possess high levels of sequence similarity. The approaches that are most commonly employed to estimate evolutionary distances between prokaryotes are based on phylogenetic analyses of rRNAs (20) or highly conserved proteins, e.g. concatenated ribosomal proteins (21). These methods generally do not provide the resolution that is required to elucidate the relationships between closely related genomes. For closely related genomes, more suitable measures of evolutionary distance are the synonymous (*dS*) and nonsynonymous (*dN*) substitution rates, ideally, for all orthologous genes. The synonymous substitution rate, which is typically one to two orders of magnitude greater than *dN*, is the more sensitive measure. Accordingly, the median of *dS* over all orthologous gene pairs was used as the intergenomic distance for the purpose of clustering. Pseudogenes were excluded from the analysis.

One of the main purposes of ATGC is to provide users with high-quality, ready-to-use multiple alignments for the set of orthologous genes from each genomic cluster. Traditionally, identification of orthologous genes is based on bidirectional best hits (BBH) but this method is prone to producing both false-positives and false-negatives, so refinement is highly desirable. The ATGCs that consist of closely related genomes allow extensive use of synteny for ortholog identification, in conjunction with BBH. Therefore, the second criterion used for building ATGCs requires that the genomes in a cluster be also alignable, that is, that a high percentage of the BBH should belong to syntenic regions, providing for reliable ortholog identification.

The procedure for the construction of the ATGCs

The rapid growth of the collection of sequenced genomes and our goal to update web resource with each new RefSeq version, requires optimization of the procedure to minimize the computation time. We used an iterative approach to build ATGCs, where the first two rounds of clustering were designed specifically to address the optimization issue.

Step 1. All available prokaryotic genomes were clustered into taxa at the class level as defined in the NCBI taxonomy. In the next round of clustering, genomes from each class were analyzed separately.

Step 2. Rough clustering at this step was based on the sequence similarity between highly conserved genes. The collection of COG (12) profiles was downloaded from the NCBI ftp site and COG identifiers were assigned to genes using the rpsblast program from the BLAST package (22). Best hits with *E*-values <0.1 were used to assign genes to COGs. A gene was considered highly conserved if the corresponding COG was represented in at least 99% of the analyzed bacterial or archaeal genomes (bacterial and archaea were treated separately). As a result, 96 bacterial and 75 archaeal COGs were identified as highly conserved and used for subsequent clustering.

For each cluster obtained on the previous step, all genes from a given highly conserved COG were pooled, a multiple alignment was built using the MUSCLE program (23,24), and pairwise amino acid distances were estimated using the PROTDIST program of the Phylip package

Downloaded from nar.oxfordjournals.org at Lawrence Berkeley Lab on April 21, 2011

(JTT evolutionary model; gamma-distributed site rates with shape parameter 1.0). Intergenomic distances were calculated as medians of amino acid distances over all highly conserved genes. The new set of genome clusters was obtained on the basis of an ultrametric genome tree that was constructed from the matrix of intergenomic distances using the KITSCH program of the Phylip package (25). The depth cutoff of 7% was empirically selected.

Step 3. In this round, more sensitive clustering based on complete sets of orthologous genes was performed. For a particular cluster from Round 2, all pairs of genomes were considered. The likely orthologous genes were identified as BBHs all-against-all BLASTP searches, and for each pair of orthologs, dS and dN were estimated from the alignments of the coding nucleotide sequences using PAML (26). The median of dS over all orthologous pairs of genes for a particular genome pair was used as the intergenomic distance. Using the obtained matrix of intergenomic distances, an ultrametric genome tree was constructed as in Round 2. The new clusters were defined by applying a cutoff of $dS = 1.5$.

Step 4. Finally, the additional criterion of gene order conservation was applied. For each pair of genomes, all BBHs detected on the previous step were tested to determine whether or not they were supported by synteny conservation. A BBH was considered to be supported by synteny if there was a high density of adjacent BBHs in its close vicinity in both genomes. Specifically, the standard dotplot for a given genome pair was constructed from the complete set of BBHs. For each BBH, the synteny support was calculated as the maximum number of other BBHs in a sliding window consisting of seven genes and including the examined BBH. The BBHs with five or more other BBHs in the neighborhood were considered to be supported by synteny.

To determine whether or not two genomes were alignable, the rearrangement distance between two genomes was calculated as

$$DY = (Nb - Ns)/Nb$$

where Nb is the total number of BBHs and Ns is the number of BBHs supported by synteny.

To generate alignable clusters, all genomes in a cluster from the previous step were considered as nodes in a graph, edges were added if DY was <0.15, and single linkage clustering was performed. Connected components of size two or greater represent ATGCs.

Construction of clusters of orthologous genes supported by synteny

These clusters were constructed separately for each ATGC. To this end, all genes from all genomes in a particular ATGC were considered as nodes in a graph, and if two genes formed a BBH supported by synteny, they were connected by an edge, and single linkage clustering was performed. The resulting connected components represent the clusters of orthologous genes supported by synteny.

The user is provided with capabilities to download different data sets based on such clusters of orthologs, for instance, a list of the corresponding protein GIs, or multiple protein and nucleotide alignments of genes. When downloading these data sets, the user can select an arbitrary subset of genomes. When a subset is selected, the program retrieves only those clusters of orthologous genes where representative of each of selected genomes are present.

The principal features of ATGCs

Several features distinguish the ATGCs from previously available types of genomic data:

- high coverage of prokaryotic genomes in both 'vertical' (include all species/clades that fit the definition of ATGC) and 'horizontal' (genome-wide inclusion of orthologous gene sets) directions;
- reliable identification of orthologs: the high degree of similarity between the genomes in ATGCs provides for extensive use of synteny in the identification of orthologs; some of the BBHs that are not supported by synteny potentially can be xenologs, that is, homologous genes acquired by horizontal gene transfer from distant organisms (9), rather than *bona fide* orthologs connected by the species tree of the given ATGC;
- availability of reliable alignments of coding sequences, which is a prerequisite for genome-wide evolutionary analysis;
- the possibility to include noncoding regions in comparative analysis ensured by the high similarity between the genomes within ATGCs and support from synteny;
- availability of a reliable species tree underlying the set of genomes in an ATGC (deeper phylogenetic relationships among prokaryotes are hard to decipher but within tight clades, the tree topology typically can be determined with relative ease); and
- the relatively small number of nucleotide substitutions and other changes between the genomes within ATGCs allows the use of parsimony-based methods of analysis (when the observed differences directly translate into the inferred evolutionary events, it is possible to forgo complicated and assumptions-laden methods that are required to estimate the number of multiple intervening events such as reverse base substitutions, and utilize the simplicity and robustness of the parsimony principle).

Taken together, these features of the ATGCs provide for robust comparative-genomic analysis for the purpose of research in microbial microevolution.

Coverage of prokaryotes and characteristics of the ATGCs

The 104 ATGCs currently cover 446 of the 865 prokaryotic genomes that are represented in the NCBI's RefSeq database. The distribution of the ATGCs among the bacterial and archaeal taxa for which abundant sequence information is available is shown in Figure 1. Figure 2 shows the distribution of the ATGCs by the number of included genomes. Not unexpectedly, the majority of the

Downloaded from nar.oxfordjournals.org at Lawrence Berkeley Lab on April 21, 2011

ATGCs consist of two or three genomes but there are several large clusters with five or more genomes; it should be anticipated that the number of large ATGCs increases fast as the system is updated with newly sequenced genomes. Table 1 shows the basic characteristics of the largest ATGCs containing more than 10 genomes. The diversity of the data, including the wide span of genome sizes, nucleotide compositions and substitution rates, provides for a variety of evolutionary analyses that can be performed on ATGCs that are most suitable for each particular task.

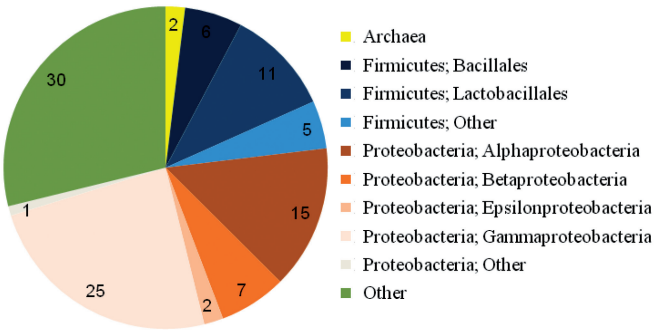

Figure 1. The distribution of the number of ATGCs among the main taxonomic groups of bacteria and archaea.

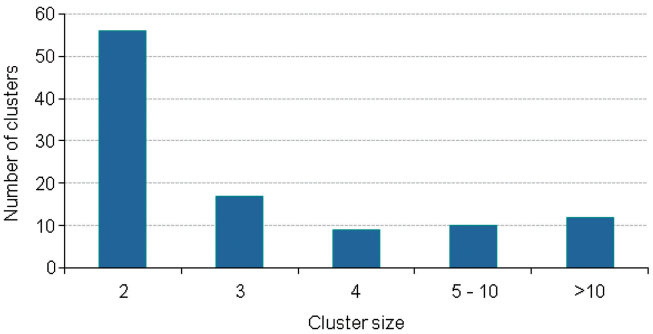

Figure 2. Distribution of the ATGCs by the number of included genomes.

Database access and interface

The ATGC website provides a highly interactive, rich interface with resizable panels, pull-down context menus, intuitive selection of genomes and navigation, and support for data sorting and filtering. The tables are easily customized, i.e. intuitive features such as adding/removing, reordering and resizing the columns, scrolling, etc. are available. The interface was built based on EXT JS framework (http://extjs.com/products/extjs), which allows for rapid component development of high performance, cross-browser and rich Internet applications.

The ATGC website provides two ways of browsing the genome clusters.

Taxonomy Explorer (Figure 3, left panel) presents all ATGCs in a tree-like structure according to their taxonomy. Not all but only the major taxa, selected manually, are used, which allows the user to simplify browsing by avoiding low-informative taxa. The number of genomes in a given ATGC is shown next to its name in square brackets. A particular cluster can be selected on a mouse click, after which the full description of cluster properties appears on the right panel.

An alternative method for browsing the ATGCs is available by selecting 'Cluster List' from the navigation menu. The clusters are listed in a table (Figure 4) that is readily customizable allowing for adding/removing the columns 'on the fly', filtering and sorting on an individual column, repositioning the columns by 'drag and drop', etc., thus providing a rich user experience, comparable with desktop applications. All ATGCs in a table are clickable which allows a user to easily navigate to the full description of a particular cluster, and simultaneously the selected cluster will be highlighted in the Taxon Explorer.

The individual cluster panel (Figure 3, right panel) shows the details on a selected ATGC. The panel contains two tables. The top table displays the phylogenetic tree along with the properties of each genome such as a genome size, GC content, links to other databases, etc. The phylogenetic tree is based on the median *dS* over all orthologous pairs of genes and provides a high resolution.

Table 1. Characteristics of large (>10 species) ATGCs

Phylum	Genus	Number of species	Genome size, Mb	GC content	*dS*	*dY*
Proteobacteria	*Escherichia*/*Shigella*/*Enterobacter*/ *Citrobacter*/*Salmonella* sp.	34	4.8 (3.0–5.5)	50.8 (50.3–56.7)	(0.00–1.99)	(0.00–0.27)
Proteobacteria	*Burkholderia* sp.	21	7.0 (5.2–7.5)	68.3 (67.6–68.5)	(0.00–0.35)	(0.00–0.14)
Proteobacteria	*Yersinia* sp.	16	4.6 (4.3–5.1)	47.6 (46.9–49.0)	(0.00–1.08)	(0.00–0.23)
Proteobacteria	*Haemophilus influenzae*	13	1.9 (1.8–2.0)	38.0 (38.0–38.2)	(0.00–0.09)	(0.00–0.13)
Firmicutes	*Staphylococcus aureus*	13	2.9 (2.7–2.9)	32.8 (32.7–32.9)	(0.00–0.05)	(0.00–0.04)
Firmicutes	*Listeria* sp.	13	3.0 (2.8–3.1)	37.8 (36.4–38.0)	(0.00–1.03)	(0.01–0.14)
Proteobacteria	*Vibrio cholerae*	13	4.0 (3.8–4.1)	47.5 (47.4–47.6)	(0.00–0.04)	(0.01–0.26)
Firmicutes	*Streptococcus pyogenes*	12	1.9 (1.8–1.9)	38.5 (38.3–38.7)	(0.00–0.02)	(0.00–0.04)
Firmicutes	*Bacillus* sp.	12	5.2 (4.1–5.6)	35.4 (34.8–35.9)	(0.00–1.82)	(0.00–0.24)
Proteobacteria	*Shewanella* sp.	11	5.0 (4.7–5.3)	46.2 (44.4–47.9)	(0.01–1.72)	(0.00–0.18)
Proteobacteria	*Campylobacter* sp.	11	1.7 (1.6–1.8)	30.5 (30.3–31.1)	(0.00–1.88)	(0.00–0.09)
Firmicutes	*Streptococcus pneumoniae*	11	2.1 (2.0–2.2)	39.7 (39.6–39.8)	(0.00–0.01)	(0.00–0.08)

Median, minimum and maximum values are given for genome size and GC content; minimum and maximum values are given for synonymous substitution rate (*dS*) and synteny distance (*dY*).

Downloaded from nar.oxfordjournals.org at Lawrence Berkeley Lab on April 21, 2011

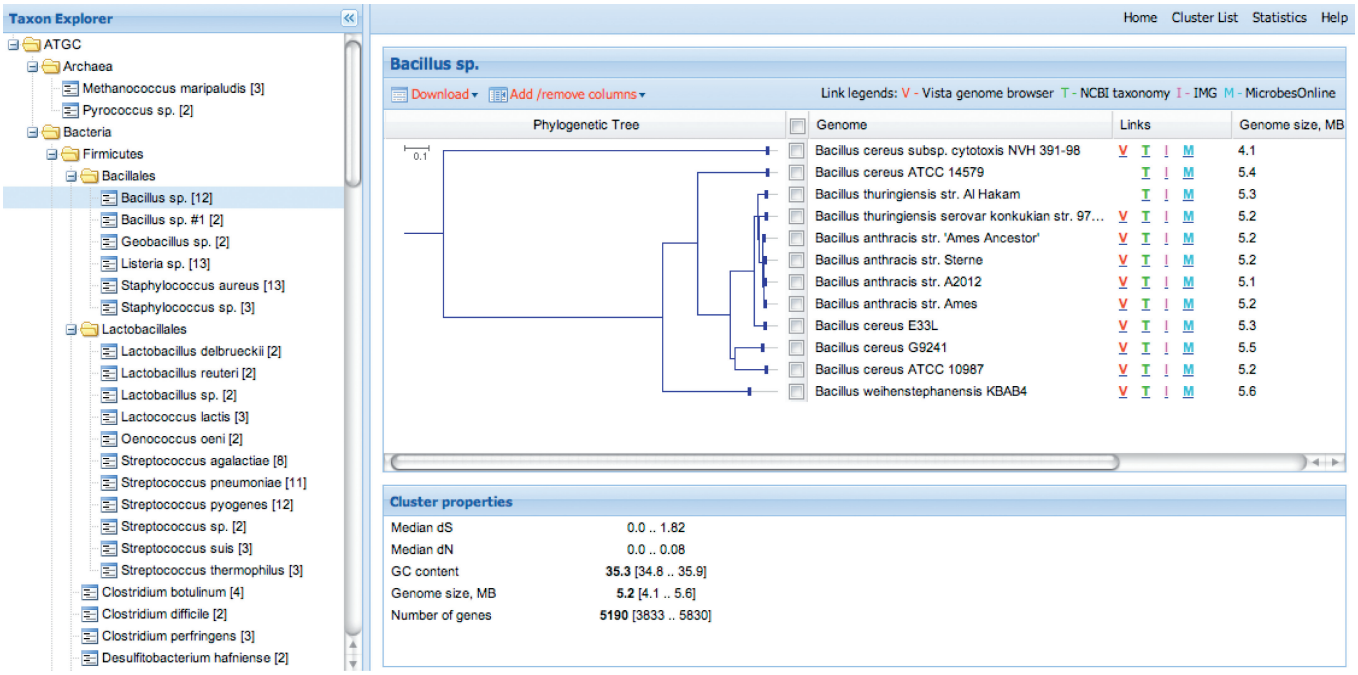

Figure 3. A screen shot of the ATGC web page. On the left panel is the taxon explorer with one of the clusters selected, and on the right panel are: site navigation menu on the top, and genomes and properties tables for the selected cluster in the bottom.

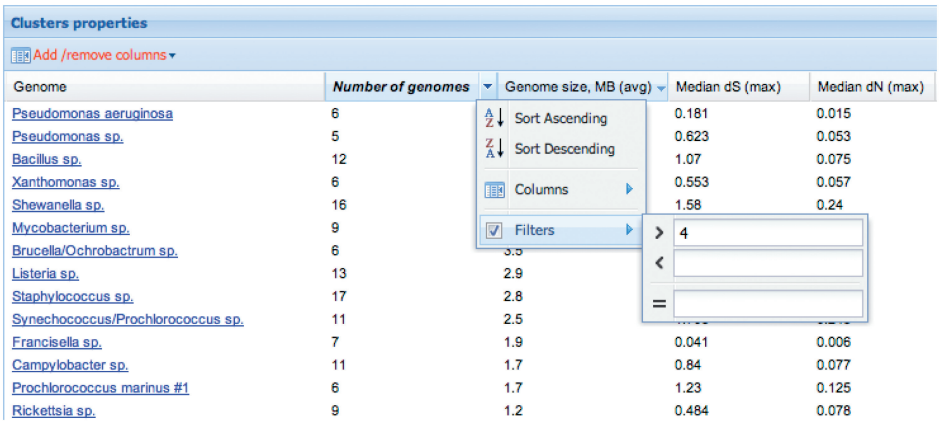

Figure 4. Clusters properties table sorted by genome size and applying a filter to a cluster size (number of genomes in a cluster). A filtered column can be then easily identified by a bold italic header.

The bottom table shows overall statistics of a cluster, such as average, minimum and maximum values of the GC content, genome size, etc.

The primary purpose of this panel is to provide users with an easy and effective way to select the genomes that have the appropriate properties for the task at hand and to download the desired data set. A user is supposed to select a set of genomes by clicking on checkboxes which gets Download pull-down menu enabled (Figure 5).

In the current version we provide the following types of data for downloading: the list of protein GI numbers for each group of orthologs supported by synteny, multiple alignments of orthologous ORFs, multiple alignments of orthologous proteins and the two matrices of pair-wise genome distances based on the median of *dN* and median of *dS*.

CONCLUSIONS

The ATGC resource exploits the rapidly accumulating genome sequences of prokaryotes to create a platform for research projects in microbial microevolution. The availability of ATGCs with widely different characteristics and representing diverse bacteria and archaea provides for the possibility to address a variety of evolutionary questions. Such questions include, for example, constancy or variability of evolutionary rates and selective pressure within clusters of closely related genomes; relationships between sequence evolution and genome rearrangements; dependence of selection processes, gene loss and acquisition and other aspects of genome evolution on the life style of bacteria and archaea; and many other problems of microbial evolution. It can be expected that the

Downloaded from nar.oxfordjournals.org at Lawrence Berkeley Lab on April 21, 2011

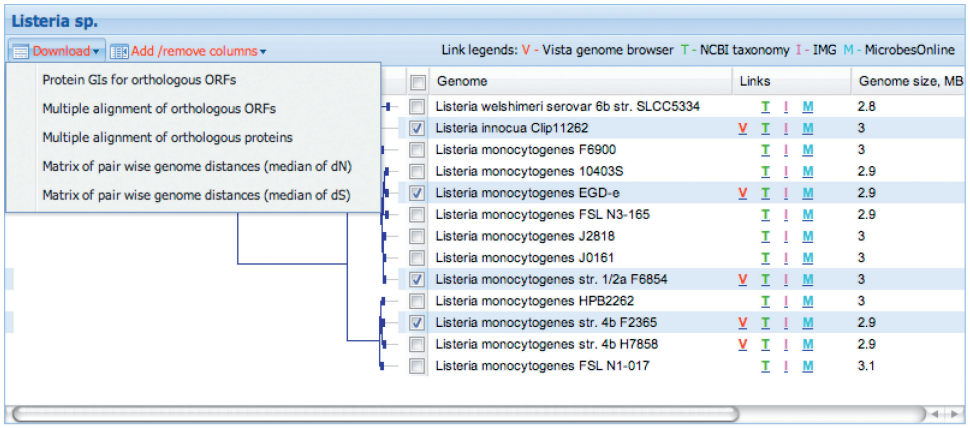

Figure 5. The 'Download pull-down menu' allows for the download of the various types of data for a selected list of genomes.

availability of ATGCs that will be continuously updated to include new genomes substantially boosts and facilitates research in comparative genomics and evolution of bacteria and archaea.

Future plans

A projected enhancement of the system will be the customization of the ATGC-building procedure so that a user will have the capability to change the *dS* and *DY* cutoffs, a feature that will allow the user to expand or contract genome clusters on the fly. We are planning to keep the ATGC database up to date by regularly incorporating genomes from new RefSeq versions.

ACKNOWLEDGEMENTS

This work was part of the Virtual Institute for Microbial Stress and Survival (http://VIMSS.lbl.gov).

FUNDING

US Department of Energy (DE-AC02-05CH11231); Department of Energy Joint Genome Institute (to I.R.); the intramural research program of the US Department of Health and Human Services (National Library of Medicine, NIH to Y.I.W. and E.V.K.).

Conflict of interest statement. None declared.

REFERENCES

1. Koonin,E.V. and Wolf,Y.I. (2008) Genomics of bacteria and archaea: the emerging generalizations after 13 years. *Nucleic Acids Res.*, **36**, 6688–6719.
2. Liolios,K., Mavromatis,K., Tavernarakis,N. and Kyrpides,N.C. (2008) The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.*, **36**, D475–D479.
3. Jordan,I.K., Kondrashov,F.A., Rogozin,I.B., Tatusov,R.L., Wolf,Y.I. and Koonin,E.V. (2001) Constant relative rate of protein evolution and detection of functional diversification among bacterial, archaeal and eukaryotic proteins. *Genome Biol.*, **2**, RESEARCH0053.
4. Rocha,E.P., Smith,J.M., Hurst,L.D., Holden,M.T., Cooper,J.E., Smith,N.H. and Feil,E.J. (2006) Comparisons of dN/dS are time dependent for closely related bacterial genomes. *J. Theor. Biol.*, **239**, 226–235.
5. Charlesworth,J. and Eyre-Walker,A. (2006) The rate of adaptive evolution in enteric bacteria. *Mol. Biol. Evol.*, **23**, 1348–1356.
6. Ochman,H., Lawrence,J.G. and Groisman,E.A. (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature*, **405**, 299–304.
7. Darling,A.E., Miklos,I. and Ragan,M.A. (2008) Dynamics of genome rearrangement in bacterial populations. *PLoS Genet.*, **4**, e1000128.
8. Mau,B., Glasner,J.D., Darling,A.E. and Perna,N.T. (2006) Genome-wide detection and analysis of homologous recombination among sequenced strains of Escherichia coli. *Genome Biol.*, **7**, R44.
9. Koonin,E.V. (2005) Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.*, **39**, 309–338.
10. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
11. Makarova,K.S., Sorokin,A.V., Novichkov,P.S., Wolf,Y.I. and Koonin,E.V. (2007) Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea. *Biol. Direct.*, **2**, 33.
12. Tatusov,R.L., Natale,D.A., Garkavtsev,I.V., Tatusova,T.A., Shankavaram,U.T., Rao,B.S., Kiryutin,B., Galperin,M.Y., Fedorova,N.D. and Koonin,E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**, 22–28.
13. Li,L., Stoeckert,C.J. Jr. and Roos,D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.*, **13**, 2178–2189.
14. Fulton,D.L., Li,Y.Y., Laird,M.R., Horsman,B.G., Roche,F.M. and Brinkman,F.S. (2006) Improving the specificity of high-throughput ortholog prediction. *BMC Bioinformatics*, **7**, 270.
15. Zmasek,C.M. and Eddy,S.R. (2002) RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics*, **3**, 14.
16. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Edgar,R., Federhen,S. *et al.* (2008) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **36**, D13–D21.
17. Alm,E.J., Huang,K.H., Price,M.N., Koche,R.P., Keller,K., Dubchak,I.L. and Arkin,A.P. (2005) The MicrobesOnline Web site for comparative genomics. *Genome Res.*, **15**, 1015–1022.
18. Markowitz,V.M., Szeto,E., Palaniappan,K., Grechkin,Y., Chu,K., Chen,I.M., Dubchak,I., Anderson,I., Lykidis,A., Mavromatis,K. *et al.* (2008) The integrated microbial genomes (IMG) system in 2007: data content and analysis tool extensions. *Nucleic Acids Res.*, **36**, D528–D533.
19. Frazer,K.A., Pachter,L., Poliakov,A., Rubin,E.M. and Dubchak,I. (2004) VISTA: computational tools for comparative genomics. *Nucleic Acids Res.*, **32**, W273–W279.

Downloaded from nar.oxfordjournals.org at Lawrence Berkeley Lab on April 21, 2011

20. DeSantis,T.Z., Hugenholtz,P., Larsen,N., Rojas,M., Brodie,E.L., Keller,K., Huber,T., Dalevi,D., Hu,P. and Andersen,G.L. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.*, **72**, 5069–5072.
21. Wolf,Y.I., Rogozin,I.B., Grishin,N.V., Tatusov,R.L. and Koonin,E.V. (2001) Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol. Biol.*, **1**, 8.
22. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
23. Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.
24. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, **32.**, 1792–1797.
25. Felsenstein,J. (1989) PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics*, **5**, 164–166.
26. Yang,Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.*, **13**, 555–556.

Downloaded from nar.oxfordjournals.org at Lawrence Berkeley Lab on April 21, 2011

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.