In Silico Biology 5 (2005) 611–615
IOS Press

Short Communication

u-Genome: A Database on Genome Design in Unicellular Genomes

Kishore Ramaji Sakharkar[a], Iti Chaturvedi[b], Vincent T.K. Chow[c], Chee Keong Kwoh[d], Pandjassarame Kangueane[b] and Meena Kishore Sakharkar[b,d,*]

[a]*National University Medical Institutes, Yong Loo Lin School of Medicine, National University of Singapore, Singapore*
[b]*Nanyang Centre for Supercomputing and Visualization, MAE, Nanyang Technological University, Singapore 639798*
[c]*Department of Microbiology, National University of Singapore, Singapore*
[d]*SCE, BioInformatics Research Centre, Nanyang Technological University, Singapore 639798*

Edited by H. Michael; received 16 September 2005; revised and accepted 22 November 2005; published 1 December 2005

ABSTRACT: Unicellular eukaryotes were among the first ones to be selected for complete genome sequencing because of the small size of their genomes and their interactions with humans and a broad range of animals and plants. Currently, ten completely sequenced unicellular genome sequences have been publicly released and as the number of available unicellular genomes increases, comparative genomics analysis within this group of organisms becomes more and more instructive. However, such an analysis is difficult to carry out without a suitable platform gathering not only the original annotations but also relevant information available in public databases or obtained by applying common bioinformatics methods. With the aim of solving these difficulties, we have developed a web-accessible database named u-Genome, the unicellular genome design database. The database is unique in featuring three datasets namely (1) orthologous proteins (2) paralogous proteins and (3) statistical distributions on exons, introns, intergenic DNA and correlations between them. A tool, *Uniview*, designed to visualize the gene structures for individual genes in the genome is also integrated. This database is of importance in understanding unicellular genome design and architecture and evolution related studies. The database is available through a web interface at http://sege.ntu.edu.sg/wester/ugenome.

KEYWORDS: Unicellular genomes, design, architecture, exon-intron lengths, paralogs, orthologs

INTRODUCTION

Biological insights through comparative genomics are now possible for a large number of prokaryotic as well as eukaryotic genomes. The information derived from complete genome sequences and particularly from their comparative analysis is used explicitly to study the biology of the microbial cells and to infer phylogenetic conclusions. Genome sequencing has been completed for ten unicellular eukaryotes and is in the finishing phase for several others. Because unicellular eukaryotes represent a phylogenetically coherent group comprised of pathogens and non-pathogens, they offer, as such, an ideal

*Corresponding author. N3-2C-113B, Nanyang Technological University, Singapore 639798. Tel.: +65 6790 5836; Fax: +65 6791 1859; E-mail: mmeena@ntu.edu.sg.

Electronic publication can be found in *In Silico Biol.* **5**, 0055 <http://www.bioinfo.de/isb/2005/05/0055/>, 1 December 2005.

set of model organisms for comparative genomic studies. However, in silico comparative genomics have been hampered by several factors: (i) the lack of a suitable platform to formulate queries in comparative genomics, (ii) inconsistencies in the annotation of the genomes, (iii) the lack of integration between annotation data and other sources of information. Few individual databases addressing the issues of introns and gene function in yeast genome exist [1,2]. However, the tools available for large-scale comparative analysis on these genomes are still lacking. Also, comprehensive information on orthologous and paralogous genes and features of intron/exon are not available. We present here u-Genome, a web-accessible database that is intended to solve these difficulties. The database is unique in featuring three datasets namely (1) orthologous proteins (2) paralogous proteins and (3) statistical distributions on exons, introns, intergenic DNA and correlations between them for ten completely sequenced unicellular genomes – *Saccharomyces cerevisiae* [3], *Plasmodium falciparum* [4], *Encephalitozoan cuniculi* [5], *Schizosaccharomyces pombe* [6], *Yarrowia lipolytica* [7], *Kluyveromyces lactis* [7], *Candida glabrata* [7], *Debaromyces hansenii* [7] and *Cryptococcus neoformans* [8] and *Eremothecium gossypii* [9]. A tool, Uniview, designed to visualize the gene structures for individual genes in the genome is also integrated. u-Genome also provides an efficient system to query, visualize, and download sequence information. Features are available to visualize the positions of introns in these datasets. Such comparative analysis of genomes on a single platform can overcome the difficulties in obtaining relevant information available from primary data sources. The purpose of u-Genome is to act as a one- stop-shop for gene and genome architecture information on completely sequenced unicellular genomes. This dataset is of importance for comparative genome analysis and provides insights to unicellular eukaryotic genome design, gene function and gene evolution.

MATERIALS AND METHODS

Dataset generation and statistical analysis

The annotated genome sequence data for *Saccharomyces cerevisiae, Plasmodium falciparum, Encephalitozoan cuniculi, Schizosaccharomyces pombe, Yarrowia lipolytica, Kluyveromyces lactis, Debaromyces hansenii, Candida glabrata, Cryptococcus neoformans* and *Eremothecium gossypii* were downloaded from NCBI (ftp://ftp.ncbi.nih.gov/genomes/). The genome files were parsed using the 'CDS' annotation in the FEATURE and the protein and nucleotide sequences were extracted [10]. The information extracted for exons and introns were stored as exon and intron datasets. The methodology for construction of the datasets is described as a flowchart online. The number of exons, introns and proteins are tabulated in Table1. Statistical analysis on exons, introns, intergenic DNA, and chromosome size are available online.

Paralogous and orthologous proteins

CD-HIT was performed on the ten genomes at 60% and respective clusters of paralogous and orthologous proteins were identified [11]. Each of the clusters was subjected to ClustalW alignment and intron positions were mapped onto the protein clusters. Alignments for the clusters are available in text and picture format online with available introns positions. These data are essential for insights on intron sliding and the process of splice site reassignment. The data is available through the web interface and is accessible via a search engine.

Table 1
Distribution of number of exons, introns, proteins, paralogous clusters for genomes in u-genome

Genome	Genome size (Mega base pair)	# of exons	# of introns	# of proteins	# of paralogous clusters
S. cerevisiae	12.1	6159	312	5866	270
S. pombe	12.4	9677	4677	5000	161
E. cuniculi	2.9	2011	15	1996	66
P. falciparum	23	12658	7391	5267	70
C. glabrata	12.3	5265	84	5181	122
D. hansenii	12.2	6674	356	6318	216
K. lactis	10.6	5457	130	5327	79
Y. lipolytica	20.5	7261	740	6521	195
E. gossypii	9.2	4946	228	4718	36
C. neoformans	20	41631	35037	6594	386

Gene visualization tool – Uniview

The easiest way to gain a quick overall understanding of gene architecture is via a visual display that allows the user to view information on the positions of introns, exons and intergenic regions. Uniview, a gene visualizer allows for gene-by-gene view for a specific genome of interest among these ten unicellular genomes. It displays information on number of exons, number of introns, introns phase, splice-site consensus for each gene. Mouse over options allow the user to zoom to the desired view of the data and to print or save the images to a file. Options are provided to BLAST the individual exons, introns and intergenic regions against the pre-indexed exon and intron datasets.

PRELIMINARY DATA ANALYSIS

The genome size for the ten unicellular genomes varies from 2.9 to 23 Mega basepair. It is interesting to see that they have much smaller genomes as compared to multi-cellular eukaryotes. The genomes of these eukaryotes involve not only fewer genes but also smaller introns and intergenic regions. Preliminary analysis on data obtained suggests that exons represent more than 50% of the sequence space on all the chromosomes. This is significantly more than human and mouse for which exons represent $< 5\%$ of sequence space on all the chromosomes [12]. Introns represent less than 1% of sequence space on all chromosomes in all the genomes except *C. neoformans* in which the sequence space represented in introns is 12%–13% on all the chromosomes. This can be explained based on the fact that a significant proportion of genes in their genomes are intronless except for *C. neoformans* which is reported to be intron-rich [8]. It is also interesting to observe that unlike multi-cellular eukaryotic genomes, these unicellular genomes have greater standard deviations around the mean exon length than the standard deviations about the mean introns lengths supporting their intronless gene architectures [12]. This is also reflected in a positive correlation of > 0.99 for total length in exons and chromosome size for all the unicellular genomes (data available online). Generally, more number of genes are found on larger chromosomes in all the genomes. The results presented thus provide a framework for understanding the nature and patterns of exon-intron length distributions, the constraints on them and their role in genome design and evolution. These results for the first time suggest on the similar genome organizations in unicellular eukaryotes at the gene structure level. Detailed analyses on data provided will provide further insights to genome design in unicellular eukaryotes.

UTILITY AND AVAILABILITY

u-Genome is a comprehensive tool for genomes related information on ten completely sequenced unicellular genomes. The database aims to serve as a convenient starting point for studying the genome architecture and design in unicellular genomes. This database is fundamental for comparative genomics and evolutionary studies in unicellular eukaryotic genomes and is available at: http://sege.ntu.edu.sg/wester/ugenome/.

We plan to update u-Genome every six months.

CAVEATS

It must be noted that the traditional gene finding algorithms treat the translation start site as the 5' boundary of the gene and there are currently no computational tools to predict the non coding first exons or non coding portions of the first exon except where the true full-length mRNA sequences are available [13–15]. As this analysis is strictly based on CDS feature in genome data, it does not take into account the first exon and is biased towards internal coding exons of the gene. Nonetheless, this platform is a step towards understanding unicellular genome design.

ACKNOWLEDGEMENTS

This work is supported by A*STAR-BMRC, Singapore Grant #03/1/22/19/242.

REFERENCES

[1] Spingola, M., Grate, L., Haussler, D. and Ares, M., jr., (1999). Genome-wide bioinformatic and molecular analysis of introns in Saccharomyces cerevisiae. RNA **5**, 221-234.
[2] Lopez, P. J. and Seraphin, B. (2000). YIDB: the Yeast Intron DataBase. Nucleic Acids Res. **28**, 85-86.
[3] Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H. and Oliver, S. G. (1996). Life with 6000 Genes. Science **274**, 546, 563-567.
[4] Gardner, M. J., et al. (2002). Genome sequence of the human malaria parasite Plasmodium falciparum. Nature **419**, 498-511.
[5] Katinka, M. D., Duprat, S., Cornillot, E., Metenier, G., Thomarat, F., Prensier, G., Barbe, V., Peyretaillade, E., Brottier, P., Wincker, P., Delbac, F., El Alaoui, H., Peyret, P., Saurin, W., Gouy, M., Weissenbach, J. and Vivares, C. P. (2001). Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi. Nature **414**, 450-453.
[6] Wood, V., et al. (2002). The genome sequence of Schizosaccharomyces pombe. Nature **415**, 871-880.
[7] Dujon, B., et al. (2004). Genome evolution in yeasts. Nature **430**, 35-44.
[8] Loftus, B. J., et al. (2005). The genome of the basidiomycetous yeast and human pathogen Cryptococcus neoformans. Science **307**, 1321-1324.
[9] Dietrich, F. S., Voegeli, S., Brachat, S., Lerch, A., Gates, K., Steiner, S., Mohr, C., Pohlmann, R., Luedi, P., Choi, S., Wing, R. A., Flavier, A., Gaffney, T. D. and Philippsen, P. (2004). The Ashbya gossypii genome as a tool for mapping the ancient Saccharomyces cerevisiae genome. Science **304**, 304-307.
[10] Sakharkar, M., Passetti, F., de Souza, J. E., Long, M. and de Souza, S. J. (2002). ExInt: an Exon Intron Database. Nucleic Acids Res. **30**, 191-194.
[11] Li, W., Jaroszewski, L. and Godzik, A. (2001). Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics **17**, 282-283.
[12] Sakharkar, M. K., Perumal, B. S., Sakharkar, K. R. and Kangueane, P. (2005). An analysis on gene architecture in human and mouse genomes. In Silico Biology **5**, 0032.
[13] Galas, D. J. (2001). Sequence interpretation. Making sense of the sequence. Science **291**, 1257-1260.

[14] Stormo, G. D. (2000). Gene-finding approaches for eukaryotes. Genome Res. **10**, 394-397.
[15] Davuluri, R. V., Grosse, I. and Zhang, M. Q. (2001). Computational identification of promoters and first exons in the human genome. Nat. Genet. **29**, 412-417.