RsGDB, the *Rhodobacter sphaeroides* Genome Database

M. Choudhary, C. Mackenzie, N. J. Mouncey and S. Kaplan*

Department of Microbiology and Molecular Genetics, University of Texas Medical School, Houston, TX 77030, USA

Received August 17, 1998; Accepted September 23, 1998

DDBJ/EMBL/GenBank accession nos+

ABSTRACT

This report provides a summary of the sequencing project of the small chromosome (CII) of *Rhodobacter sphaeroides* 2.4.1^{T}, and introduces the first version of the genome database of this bacterium. The database organizes and describes diverse sets of biological information. The main role of the *R.sphaeroides* genome database (RsGDB) is to provide public access to the collected genomic information for *R.sphaeroides* via the World-Wide Web at http://utmmg.med.uth.tmc.edu/sphaeroides . The database allows the user access to hundreds of low redundancy *R.sphaeroides* sequences for further database searching, a summary of our current search results, and other allied information pertaining to this bacterium.

INTRODUCTION

Rhodobacter sphaeroides 2.4.1^{T} is a facultative photoheterotroph belonging to the α-3 subdivision of the *Proteobacteria* (1). Members of the *R.sphaeroides* exhibit complex genomic organization, variation of genome sizes, chromosome number and metabolic versatility. Our laboratory was the first to report that *R.sphaeroides* 2.4.1^{T} possesses two different circular chromosomes (2,3), of sizes ~3.0 Mb (CI) and ~0.9 Mb (CII), which was in direct contradiction to the long held central dogma pertaining to prokaryotic genome structure, consisting of only a single circular chromosome. The existence of multiple chromosmes in some other members of this group as well as organisms outside of this group (4,5) established the fact that genomic complexity is of widespread occurrence among the eubacteria.

In order to further understand genome complexity in this bacterium and as a result of funding from the Clayton Foundation for Research, we began the *R.sphaeroides* genome project with the objective of providing a low resolution DNA sequence analysis of CII and releasing the genetic and molecular information obtained to the public. Partial results of the sequencing project of CII have been published elswhere (6), but this paper provides a summary of the sequencing project of CII and integrates the genome database describing the physical map (7), partially ordered cosmid collection (4), DNA sequences mapped to both chromosomes, and allied biological information pertaining to this bacterium. Separate tables provide a list of gene matches and their putative functions, as well as a list of several hundred Bluescript subclones from which the DNA sequences were obtained. A large literature collection on *R.sphaeroides*, and contact information for principal investigators in the field are also provided. The database is fully operational and easily accessible with BLASTX similarity searching capability.

RESULTS UPDATE: SEQUENCE ANALYSIS OF CII

An updated summary of the DNA sequence data is described in Table 1. We utilized both partial and complete random sequencing strategies (5) to obtain DNA sequences from 42 cosmids containing ~757 kb of unique insert DNA of chromosome II. A total of 880 subclones were sequenced from their ends, and subsequently 1758 sequence fragments were assembled to give rise to 688 sequence contigs representing a total of ~533 kb unique DNA sequence. The average gap length was estimated to be ~325 nt. In a number of cases where we sequenced genes of interest, we filled the sequence gaps using the primer walking method (8). The total unique DNA sequence covers ~70% of the total cosmid insert DNA and about half of CII.

Analysis of unique DNA sequences resulted in 266 putative ORF-encoding genes representing a wide variety of functions, two rRNA operons, and eight tRNA genes. A partial list of these genes was reported in our earlier publication (4), and in this paper we include all genes, both newly found and previously published.

DATABASE DESCRIPTION

RsGDB, the *Rhodobacter sphaeroides* Genome Database, located at The University of Texas Health Science Center at Houston, provides access to all formats of the database, including the online DNA sequence similarity searches. This database can be accessed at the World-Wide Web address http://utmmg.med.uth.tmc.edu/sphaeroides

The database organizes the genomic information of *R.sphaeroides*, and describes the genome features, the physical map, partially ordered cosmid collections of both chromosomes (CI and CII), and DNA sequences of both CI and CII which were either previously sequenced and deposited into the GenBank, and new sequences obtained for CII in our laboratory. The physical map outlines various restriction fragment sizes of the genome, produced by rare cutting restriction enzymes. Two other tables are also provided, one with a list of all gene matches and their respective gene functions, and another with a list of Bluescript subclones along with other related information, such as insert sizes, and potential gene(s) located in that subclone.

*To whom correspondence should be addressed. Tel: +1 713 500 5502; Fax: +1 713 500 5499; Email: skaplan@utmmg.med.uth.tmc.edu

+AQ012082–AQ012213 and B07699–B07848

Table 1. Summary of CII sequencing project

Size of CII	910 kb
Number of cosmids sequenced	42
Total unique insert DNA	757 kb
Double-stranded templates	880
Average edited read length (% ambiguous bases)	580 ± 164 (2.3%)
Total DNA sequenced	947 kb
Total cosmid insert DNA sequence (includes multiple clones and overlaps)	940 kb
Fold coverage of the total cosmid inserts	1.7×
Total unique DNA sequence	533 kb
Unique sequence coverage	0.7×
mole G+C%	65.3
DNA resulting in significant database matches	222 kb
DNA resulting in the designation of putative orfs (pORFs) with no database matches	229 kb
DNA resulting in neither an orf nor a database match	82 kb
Number of sequence fragments in random assembly	1758
Number of contigs	688
Average gap length	325 bp
Number of presumptive genes identified	264
Number of ribosomal RNA operons identified	2 (*rrnB* and *rrnC*)
Number of tRNA genes identified	8 (2 Ala, Ile, 2 f-Met, 2 Met, and Val)

In addition, a comprehensive list of bibliographical references, stock list, and investigators directory are also included. RsGDB maintains a comprehensive bibliography of all journal publications and reviews. Currently we have over 1000 references in the database. RsGDB allows investigators to use the stock lists from our laboratory and intends to include the list of strains of other individual laboratories, if these are provided. RsGDB also keeps a directory of principal investigators and further information on their research interests. The detailed information about the investigators may be obtained through their linked web sites, and they may be contacted via their email addresses in the database. The listing of the investigators who ever researched with '*sphaeroides'* is not yet complete. Investigators are requested to send their names and addresses so that they can be listed on this site. All these datasets are to be updated every six months.

SEQUENCE SIMILARITY SEARCH AND BLAST RESULTS

RsGDB will provide an access to the genomic sequences of *R.sphaeroides* and a web BLAST service for matches only against *R.sphaeroides* sequences. These sequences include: all sequences from GenBank that include the word '*sphaeroides'* as an organism and the Bluescript subclone end sequences which we recently obtained under the sequence skimming project of CII in our laboratory. All the redundent sequence files from the sequence collections have been removed and all detected overlapping regions have been merged. At present more than 700 end sequences are available in the database and as we obtain any new sequences, these will be made available. The sequence files are organized as text files in fasta format. This format is very similar and recognized by many sequence analysis packages and retrieval systems. An input page for a similarity search allows one to make simple selections (blasn or blastx), and to submit the query sequence by pasting it in the box. These end sequences are also useful to researchers because they can be used to identify a subclone and a cosmid of interest.

Nucleotide sequence accession numbers

The DNA sequences described in the summary results were entered as GSS files into GenBank (NCBI). Their accesssion numbers are AQ012082–AQ012213 and B07699–B07848.

CONTACT

Queries and any correspondence pertaining to this database information including DNA sequences, cosmids and Bluescript subclones should be directed to Dr Samuel Kaplan at skaplan@utmmg.med.uth.tmc.edu. Users are invited to suggest any new sets of data they would like to see on this site, they are also invited to provide feedback regarding this site and are requested to cite this article and acknowledge this database when the database has been helpful in preparing their publications.

ACKNOWLEDGEMENTS

We would like to thank Dearl Neal for helping establish our www server. Partial results from our sequencing project are described in *Microbiology* (6) (October issue, 1997). We are also grateful to all colleagues who provided their sequence data and additional information for this publication. This work was entirely supported by a grant from the Clayton Foundation for Research.

REFERENCES

1 Woese,C.R., Stackebrandt,E., Weisburg,W.G., Paster,B.J., Madigan,M.T., Fowler,V.J., Hahn,C.M., Blanz,P., Gupta,R., Nealson,K.H. and Fox,G.E. (1984) *Syst. Appl. Microbiol*., **5**, 315–326.
2 Suwanto,A. and Kaplan,S. (1989) *J. Bacteriol*., **171**, 5850–5859.
3 Suwanto,A. and Kaplan,S. (1992) *J. Bacteriol*., **174**, 1135–1145.
4 Choudhary,M., Mackenzie,C., Nereng,K., Sodergren,E., Weinstock,G.M. and Kaplan,S. (1994) *J. Bacteriol*., **176**, 7694–7702.
5 Jumas-Bilak,E., Michaux-Charachon,S., Bourg,G., Ramuz,M. and Allardet-Servent,A. (1998) *J. Bacteriol*., **180**, 2749–2755.
6 Choudhary,M., Mackenzie,C., Nereng,K., Sodergren,E., Weinstock,G.M. and Kaplan,S. (1997) *Microbiology*, **143**, 3085–3099.
7 Suwanto,A. and Kaplan,S. (1989) *J. Bacteriol*., **171**, 5840–5849.
8 Mouncey,N.J., Choudhary,M. and Kaplan,S. (1997) *J. Bacteriol*., **179**, 7617–7624.