Final Report
June 5, 2002

Rhodopseudomonas palustris **genome Workshop to be held in Spring of 2001**
Award # DE-FG02-01ER63102

Caroline S. Harwood , PI
University of Iowa, Iowa City, IA 53342

Daniel W. Drell , Program Manager.

The *Rhodopseudomonas palustris* genome workshop took place in Iowa City on April 6-8, 2001. Dr. Frank Larimer from ORNL had previously done a computational annotation of the *R. palustris* genome. The purpose of the meeting was to instruct members of the annotation working group in approaches to accomplishing the "human" phase of the *R. palustris* genome annotation. The annotation group is comprised of Dr. Frank Larimer (ORNL), Drs. F. Robert Tabita and Janet Gibson (Ohio State University), Dr. Jane Gibson (Cornell University), Dr. Thomas Beatty (University of British Columbia) and Dr. Caroline S. Harwood (University of Iowa). Graduate students and postdoctoral fellows, who are members of the Tabita, Beatty and Harwood laboratories also attended the meeting.

The goals of the meeting and of the human phase of the *R. palustris* annotation project in general were: 1) To help Frank Larimer get the *R. palustris* genome to the point where it can be submitted to GenBank. 2) To identify major points that should go into a paper describing the *Rhodopseudomonas* genome. This requires that we synthesize the immense amount of information afforded by the genome to the point where we can develop a view of what it is that makes *Rhodopseudomonas palustris* who it is.

By way of background, *Rhodopseudomonas palustris* has one large circular chromosome that is 5.49 Mb in size. Scientists at the Joint Genome Institute led by Patrick Chain did a wonderful job of assembling the genome, so we can be certain about the position of each gene relative to every other gene.

About six weeks before the meeting Frank Larimer provided the annotation team with an extensive computational analysis of the *R. palustris* genome. This included gene models as predicted with three different gene caller programs, the results of searches of each deduced gene product against four different databases (KEGG, GenBank, COGS and Pfam) and a computationally assigned function for each gene.

This information was presented to us as a contig-level viewer on the Web (with accompanying "gene edit" pages, see below) and as a "feature table". The feature table allows one to view the *R. palustris* genome in graphical form. Each of 4,820 genes in the genome is displayed in its correct orientation on the chromosome. The complete nucleotide sequence of the genome can be viewed, nucleotide by nucleotide. Each gene can be viewed at the nucleotide and amino acid level. One can scroll though the genome, click on a particular gene and view, in detail, all elements of the computational annotation provided by Dr. Larimer.

In preparation for the meeting each lab in the annotation group analyzed about a 1 Mb segment, around 1,000 genes. At least one person reviewed each gene in the genome to confirm the gene modeling calls and the gene identifications, and where necessary, to extend the categorization of each gene. For each gene we asked ourselves, is the computational assignment likely to be correct? Should a more precise or different assignment be made? We also identified places in the genome where genes might need to be added. For example, a team member may notice a gap of 2 or 3 kb in the genome that does not contain a "called" gene. It

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or 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 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. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

may be that on close inspection it is apparent that there is a gene present that needs to be added to the gene inventory.

Human changes to the computational annotation were made on "edit" pages that Dr. Larimer provided for each gene. In the end it is the information on each of these gene edit pages that will be submitted to Genbank by Dr. Larimer.

Prior to the meeting each group also took an in-depth look at their favorite genes and at genes that define the human view of *R. palustris*. These included carbon dioxide fixation genes, sulfur metabolism genes, hydrogenase genes, nitrogenase genes, and photosynthesis genes.

<u>During the annotation meeting we focused on four areas.</u>

1) We spent a lot of time discussing issues related to not very glamorous aspects of the genome. For example, gene modeling: how can one be sure that the start site that was computationally called is correct, and how does one recognize and annotate possible frameshifts? The issue of how to identify missing genes, not picked up by the computer as being real, was discussed. This was time well spent. Dr. Larimer did a tremendous job of educating us about genomics and the annotation process and we left the meeting realizing that - although we would each have to "reannotate"' many of the genes that we 'd annotated before the meeting - this time we'd get it right and be consistent!

2) We pooled our knowledge about biological properties of *R. palustris* and started to relate this knowledge to the genome. This helped us to start to develop a picture of who *Rhodopseudomonas palustris* is. Based on anecdotal evidence, *R. palustris* may well be one of the most abundant species of bacteria on earth. It can be found in virtually any temperate soil or water sample that one cares to check. It is a very robust bacterium that is able to survive for long periods of time with very few nutrients. Finally, *Rhodopseudomonas palustris* is one of the most metabolically versatile bacteria known. Each of these aspects of its biology is reflected by its genome. *R. palustris* has genes for the catabolism of diverse kinds of carbon sources, including lignin monomers, fatty acids and dicarboxylic acids. It encodes two different carbon dioxide fixation enzymes and three different nitrogen fixation enzymes, each with a different transition metal at its active site. It has genes to carry out anaerobic respiration using nitric oxide and nitrite as electron acceptors and it has genes for thiosulfate oxidation. It is obvious that an organism with this degree of metabolic versatility must have a lot of traffic with its environment. This is reflected by a very large number of transport systems, especially transport systems for iron. Genes for at least seven multidrug resistance drug efflux pumps and the presence of a cluster of genes for polyketide biosynthesis may help explain why *R. palustris* survives so well in most soil and water environments.

3) In keeping with other microbial genomes, about 20% of *R. palustris* genes are of unknown function and found only in *R. palustris* and another 20% of the *R. palustris genes* are homologous to genes of unknown function found in other organisms. The annotation group realized that when one reviews the genome on a gene by gene basis one frequently sees that the "best hit" is to a gene from one of three bacteria: *Rhodobacter sphaeroides*, *Caulobacter crescentus* or *Mesorhizobium loti*. This prompted us to ask Dr. Larimer to do more work. Since the annotation meeting, Frank Larimer has done the following. a) Each *R. palustris* gene has been compared to each *Rhodobacter sphaeroides* gene. This comparison has allowed us to focus on those genes that are shared by two species of purple nonsulfur phototrophic bacteria. Although these two species have some characteristics in common metabolically, they are not extremely closely related phylogentically. b) Each *R. palustris* gene has been compared to each *Mesorhizobium loti* gene. 16S rRNA analysis indicates that *R. palustris* is very closely related phylogentically to *Mesorhizobium loti*. Does a close look at genes shared by a nitrogen fixing plant symbiont and *R. palustris* help us better understand what it is that makes *R. palustris* who it is? c) Each *R. palustris gene* has been compared to each *Caulobacter crescentus* gene. Both of these species undergoes cellular differentiation by budding cell division. Among the genes that these two species have in common, can we identify a set of genes that may be involved in development?

4) The annotation group used genomic data to start doing a metabolic reconstruction of *R. palustris*. A few days prior to the meeting in April, a software engineer at ORNL wrote programming that allowed the *R. palustris* genome to be "mapped' to the KEGG metabolic maps. It is now possible to look at metabolic processes such as lipid metabolism, cofactor and vitamin biosynthesis, or energy metabolism and see if *R. palustris* has the genes needed to carry out the core set of metabolic functions that is indicated by the maps. It is also now much easier to determine the extent to which *R. palustris* may have redundant genes for a given function or alternative routes for a particular metabolic function.

Since the meeting the annotation group has hand edited and, where necessary, reannotated each of the approximately 4,820 genes in the *Rhodopseudomonas* genome. We have also taken a close look at each of the *R. palustris* metabolic maps. The purpose of this was to see if everything that we think should be there in order for *R. palustris* to accomplish a particular metabolic function IS there. Finally, we continue to communicate by e-mail to identify aspects of the *Rhodopseudomonas palustris* genome that we should look at in depth to get a better idea of this bacterium's biological potential and identity.

A partial draft of a paper describing the *Rhodopseudomonas palustris* genome has been written and a full version of the paper should be ready for submission by the end of the summer 2002.