Annotation of the *Nanoarchaeum equitans Kin4-M* Genome from DNA Coordinates 50213 to 54031 (or Locus Tags NEQ057 to NEQ059)

Lucy Michael, Giana Milazzo, Sydney McKinnon, and Peter Hentschke
The Harley School, Brighton NY and the Western New York Genetics in Research and Health Care Partnership

SEPA SCIENCE EDUCATION PARTNERSHIP AWARD
Supported by the National Institutes of Health

Abstract

Three genes from the Archaeum *Nanoarchaeum Equitans* (NEQ057..NEQ059) were annotated using the collaborative genome annotation website geni-act. Geni-act does not have a proposed gene product prediction for these genes. Each was assessed in terms of the general genomic information, amino acid sequence-based similarity data, structure-based evidence from the amino acid sequence, and cellular localization data. Based on the analysis, proposed annotations for the gene products were determined.

Introduction

Nanoarchaeum equitans was discovered in 2002 and is a member of the marine Archaea species. It has been proposed as the first species of a new phylum. This organism grows best in environments with a pH of 6 and a salinity concentration of 2%. Traces of this microbe have also been found on the Sub-polar mid Oceanic Ridge as well as in the Obsidian Pool in Yellowstone National Park. (Huber, et al., 2002)

The genome consists of one circular chromosome with an average GC-content of 31.6%. It doesn't have most genes required for synthesis of amino acids, nucleotides, cofactors, and lipids. The fact that it belongs in the Archaea domain was determined after examination of single-stranded RNA. However, after closer examination, it was determined that it is its own phylum, which scientists/researchers named *Nanoarchaeum*. This genome lacks the ability to metabolize hydrogen and sulfur. It does have five subunits of ATP synthase, but it's currently unknown whether it obtains energy from biological molecules imported from *Ignicoccus* or if it receives energy directly. (Huber, et al., 2002)

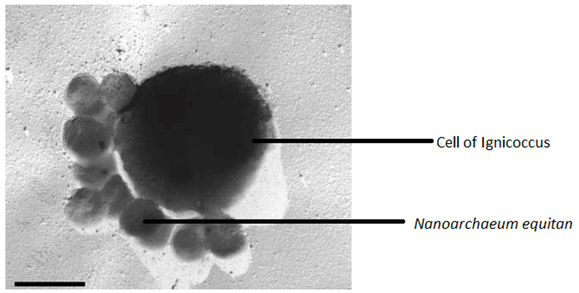

Figure 1. Image of *Nanoarchaeum equitans* species (Huber et. al., 2002)

Methods

Modules of the GENI-ACT website (http://www.geni-act.org/) were used to complete portions of the *Nanoarchaeum Equitans Kin4-M* genome annotation. The modules are described below:

Modules	Activities	Questions Investigated
Basic Information	DNA Coordinates and Sequence, Protein Sequence	What is the sequence of the gene and protein? Where is it located in the genome?
Sequence-Based Similarity	Blast, CDD, T-Coffee, WebLogo	How similar is the protein under investigation to other proteins in GenBank?
Structure-Based Similarity	TIGRfam, Pfam, PDB	What functional domains are present in the protein under investigation?
Cellular Localization	Gram Stain, TMHMM, SignalP, LipoP, Psortb, Phobius	Is the protein under investigation located in the cytoplasm, secreted, in the periplasm or embedded in the cell membrane or cell wall?
Final Annotation	Evaluate data from all modules	Has the gene been correctly called by the pipeline annotation?

Results

NEQ057:

NEQ057 has the DNA coordinates 50213...51247 and the sequence length was 1035 bases. NEQ057's two top BLAST hits were both cell division control protein Cdc6's from: (1st hit) *Desulfurococcales archaeon*, and (2nd hit) *Candidatus bathyarchaeota archaeon*. In PSORT-b, it was predicted that the protein would be found in the cytoplasm. In TMHMM, no evidence of transmembrane helices were found. Lastly, while using Phobius, it was shown that there were no signal peptides on the protein. All of this information backs the inference that this protein is located in the cytoplasm.

In the CDD search, COG, Pfam and TIGR hits were found. The TIGRFAM hit was named orc1/cdc6 family replication initiation protein and had the number 02928. The Pfam hit was named CDC6, C terminal winged helix domain and had the number pfam09079. It was part of the clan HTH CL0123, which is a family that contains a wide range of mostly DNA-binding domains that have a helix-turn-helix motif.

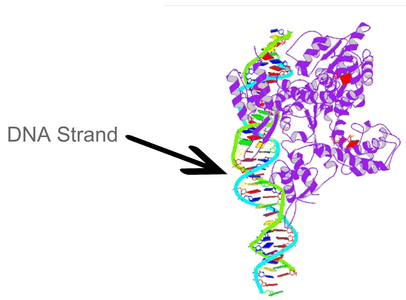

Figure 2. Proposed crystal structure of gene NEQ057, with DNA strands bound to it.

NEQ058:

NEQ058 has the DNA coordinates 51760..52191 as given by Geni-Act. The DNA sequence is 432 bases long and included 143 amino acids. The top two BLAST hits were for a 30S ribosomal protein *[Candidatus Nanobsidianus stetteri]* and 30 S ribosomal protein S12 [*Candidatus Nanopusillus acidilobi].* TMHMM found no transmembrane helices and SignalP did not detect any signal peptides in the gene. PSORT-b predicted the protein would be found in the cytoplasm since the cytoplasm score was the highest. The Phobius analysis confirmed the prior results. This evidence supports that this protein is located in the cytoplasm and this is consistent with a ribosome protein.

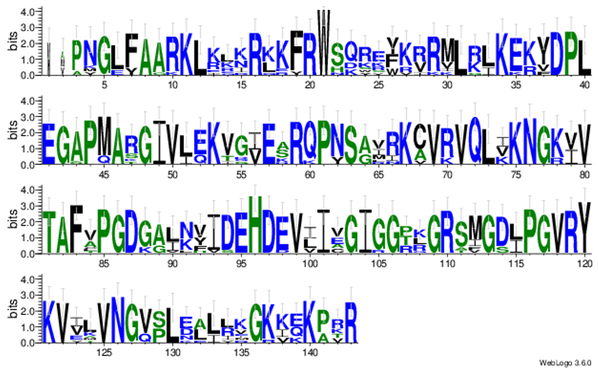

Figure 3 . NEQ058 WebLogo NR results. The WebLogo is fairly well conserved across the whole sequence. The amino acids are colored according to their chemical properties. The majority of the amino acids are polar (shown in green) or basic (shown in blue).

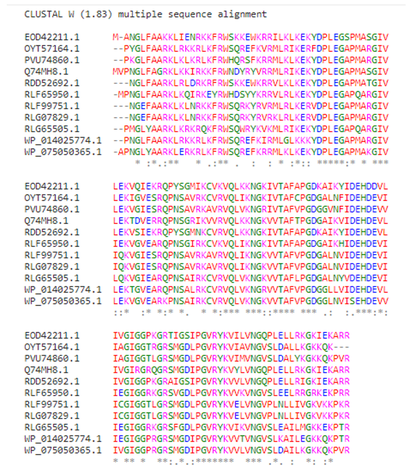

Figure 4. The multiple sequence alignment of NEQ058 in T-Coffee.

NEQ059:

The gene product given by the top BLAST hit was an ATPase from the *Candidatus Woesearchaeota archaeon* genome. The alignment length is 607, the score was 584 bits, and the E-value given was 0.0. NEQ059 has the DNA coordinates of 52238...54031, and the base sequence length was 1794.

TMHMM predicted that there were no transmembrane helices. Then in SignalP, no signal peptides were found to be on the protein. PSORT-b predicted that the protein would be found in the cytoplasm. This further backs the hypothesis that the protein is found in the cytoplasm.

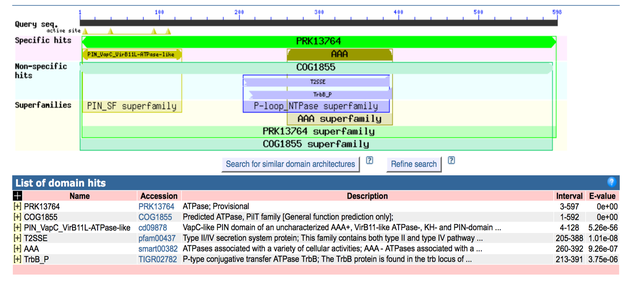

Figure 5. NEQ059 CDD results. The CDD search came up with multiple hits, including COG, Pfam and TIGR hits. The TIGRFAM hit was named TrbB with the number 02782. The top significant hit for Pfam was named T2SSE and was part of the P-loop NTPase clan. All of these domain matches have some type of ATPase activity.

Conclusion

Based on the analysis of the modules described above, the proposed gene product annotations of these genes are listed below.

Locus Tag	Proposed Annotation
NEQ057	**ORC complex protein Cdc6/Orc1**
NEQ058	**SSU ribosomal protein S12P**
NEQ059	**ATPase, PilT family**

References

Huber et al., 2002. Nanoarchaeota: New life under the sea. The Genome News Network.

University at Buffalo The State University of New York
Jacobs School of Medicine and Biomedical Sciences
buffalo.edu

Acknowledgments

Supported by an NIH Science Education Partnership (SEPA) Award - R25GM129209