DATABASES AND INFORMATION RETRIEVAL

In this exercise we will focus on getting together information on the archaeon "*Cenarchaeum symbiosum*" (Csymb) from a variety of biological data resources, because we have a project starting on Csymb proteomics. As good scientists, we first want to know who Csymb is, who is it related to, where does it live, what does it do, does it have its genome sequenced etc…

The resources we will use range from primary sequence databases, to microbial nomenclatural resources; therefore, at the end of this exercise you should have a good idea of where to get taxonomic, phylogenetic, DNA/protein sequence, metabolic, and genomic data.

1. Who is *Cenarchaeum symbiosum*?

In the first part of the exercise we will gather information on our archaeon's identity, relatives, and general phenotype, growth characteristics etc…

Start browsing the "List of Prokaryotic Species with Standing in Nomenclature" (LPSN) website at:

http://www.bacterio.cict.fr/

This is an excellent hand-made resource for all your needs in cultured bacterial and archaeal species; current name, original publication, or classification.

Search this website to answer the following questions:

- *Full taxonomic classification of Csymb (domain, phylum, class…)*

- What does the term "without standing in nomenclature" mean? Does Csymb have standing in nomenclature?

- Where was the sample that the scientists isolated Csymb obtained?

- What is the hypothesized temperature adaptation of Csymb?

- What are the isolation sources for relatives (from a sequence identity perspective) of Csymb?

- Has the classification of Csymb change from what is proposed in the original publication?

- What is the INSDC sequence accession number of the 16S rRNA sequence from Csymb?

If you have the INSDC sequence accession number, it is time to move to sequence-centric resources. We'll start with a secondary database, SILVA ribosomal RNA database, and try to get some data on relatives of Csymb. Go to:

http://www.arb-silva.de/

Here, use the search page to find the sequence information that SILVA provides on Csymb, and try to answer the following questions:

- What is the sequence quality according to SILVA?

- Is this sequence likely to be a chimera (check the definition of chimeric sequences and the pintail program)?

- *What is the sequence length?*

Once you are done with the above questions, follow the link to ENA and download the Csymb sequence in FASTA format. We will now use the "search" function of the "SINA" online aligner on SILVA website to retrieve sequences related to Csymb sequence. Find the aligner page, upload the FASTA sequence that downloaded from ENA, and set the following parameters for the aligner:

- Check the box "search and classify"

- Set "min identity with query sequence" to 0.85

- Set "number of neighbors to query sequence" to 50

- Under "advanced alignment parameters" check the option saying "Archaea" under the "Variability profile" menu

- Under "advanced search and classification parameters" change the "Ref NR" selection to "Ref"

Run the aligner and once finished, "add the neighbors to the cart". Now find the "show cart" option and start searching in your cart.

- *How many of the neighbor sequences as from "cultured" (or names organisms)?*

- Are there any rRNA sequences from genomes among these (hint: use strain field in search, or check the accession number)? If you find one, note this down.

Once you are done with answering these questions, download your cart sequences in arb file format. Normally, the next step would be to construct a phylogenetic tree with your sequence of interest and the neighboring sequences. Unfortunately, we will not have time for this.

2. Where does *Cenarchaeum symbiosum* live?

Now we will try to see where Csymb lives (or the relatives of it). We will first visit a portal called MG-RAST:

http://metagenomics.anl.gov/

Again, have a browse around; try to figure out what MG-RAST is all about and what it does. Once you are done, find the search page, and do an advanced search on organism name, with the condition that it "equals to" "Cenarchaeum symbiosum"

- How many metagenomes did you find that contain Csymb?

- Did you find Csymb in environments other than the marine environment?

- Find a couple of metagenome projects that have sequence counts larger than 500,000, and that also have a NCBI project ID number. Note this project ID numbers down

3. What does *Cenarchaeum symbiosum* do?

Now we will go back to primary databases and try to find some information on Csymb's genome and functional repertoire:

http://www.ncbi.nih.gov

Try searching for the full name of Csymb on the NCBI homepage. The search results are displayed in a combined portal called "Entrez".

- How many nucleotide, protein and genomes sequences did you find?

- Are there any other genome/metagenome projects for Csymb other than the one on the "Genomes" page?

- What is the genome size, GC content, gene content, and protein content of Csymb?

We are interested in nitrite reduction in Csymb. First find out the name of the key enzyme for nitrite to ammonia reduction by exploring the pathways in KEGG pathways:

http://www.genome.jp/kegg/pathway.html

Then search for the enzyme name in the protein details for Csymb genome.

- *Is it possible that Cysmb is a nitrite reducer?*

- *Can Csymb also reduce nitrate to nitrite?*

As a final step, we want to collect all proteins for the nitrite reduction, and therefore we will visit the UniProt resource:

http://www.uniprot.org/

Here, search for the protein (enzyme) name, and download the results as a fasta file.

- *What is the difference between "reviewed" and "unreviewed" proteins?*

- *Does the enzyme occur in eukaryotes?*

- *Which other bacterial/archaeal phyla do you find the enzyme in?*