JAMES RENNIE BEQUEST

REPORT ON PROJECT

Project Title: ARB workshop

Travel Dates: 29 May – 3 June

Location: Max Planck Institute for Marine Microbiology, Bremen, Germany

Group Member(s): Ziad El-Hajj

Aims: To attend a four full day training in the theory of phylogenetic analysis and probe design, in installing Linux and ARB, and in using the functions of ARB for sequence alignment, phylogenetic tree reconstruction and probe design.

OUTCOME (not less than 300 words):-

As part of my PhD project I am screening a marine metagenomic library for novel polysaccharides. My research is funded in part by the Leverhulme Trust, a charity that funds research into the development of novel materials, and the University of Edinburgh Centre for Science at Extreme Conditions (CSEC), a multidisciplinary centre designed to promote study of materials under extremes of temperature and pressure. The goal of the collaboration with CSEC is to ultimately isolate novel products from extremophile microorganisms in the deep sea, which have adapted to survive under high pressures and cold temperatures.

In addition to isolating novel materials, I want to gain insights into the microbial diversity present in this library. To this end, I sequenced the 16S ribosomal RNA (rRNA) genes from clones in the library, and wanted to use the ARB software package to align these sequences and construct a phylogenetic tree. Unlike other sequence alignment software, whose analysis is limited to rRNA primary structure, ARB can also take into account the secondary structure of the rRNA and provides a comprehensive package of interacting tools controlled by a single user interface. However, the software is complex and difficult to install, configure and use. The workshop is offered by the current developers of the software to help users fully take advantage of the software package.

Each of the four days of the workshop was divided into three sessions: lectures in the morning, computer demonstrations in the afternoon, and computer work in the evening. The lectures were very useful to me because of my limited knowledge of phylogenetic models prior to attending the workshop. These lectures covered various topics to help in selecting which options to use in ARB, inclding the strengths and weaknesses of online sequence databases, models of evolution, methods of tree reconstruction and bootstrapping. The evening computer work session allowed us to experiment with the skills acquired during the demonstrations using our own sequence datasets.

ARB is developed under the Linux operating system, which provides a powerful and stable environment for the calculations required by the software. The first day was devoted almost exclusively to installing and configuring Linux and ARB. Configuring ARB requires some knowledge of the UNIX shell prompt, and all necessary information was provided to us.

Demonstrations in the next 3 days were aimed at teaching us how to use ARB to align 16S rRNA sequences.

ARB can import 16S rRNA sequences in a variety of formats, including FASTA. Even if a sequence format isn't natively supported by ARB, filters can be programmed to recognise any number of formats. Sequences can be imported either to create a *de novo* alignment or into an already existing sequence database. Imported sequences can be automatically aligned with the ARB Fast Aligner, which will search for the most similar reference sequence in the Positional Tree (PT) server. The PT server represents an indexed format of the database and is necessary for faster sequence searches in ARB. Even with the Fast Aligner, manual refinement of the alignment is necessary, though it is made much easier if the Fast Aligner is used first and a closely related sequence in the tree is used as reference. One of ARB's major strengths is that it takes into account the secondary structure of the rRNA in the automatic alignment, and provides a Secondary Structure Editor to help manually refine the alignment as accurately as possible. The Secondary Structure Editor can even take into account pseudoknots, or "tertiary-like" structures, such as two loops base pairing with each other.

Once the sequences are aligned, they can be inserted into an existing tree using the ARB-parsimony tool. The resulting tree can be optimised using tree modifications that further improve parsimony. Alternatively, a new tree can be created from scratch using either the distance matrix model, maximum parsimony analysis or maximum likelihood analysis, all of which are handled via built-in tools in ARB. A variety of masks and filters are available for fine tuning the inclusion or exclusion of alignment positions, and custom filters can be created.

Although ARB was initially designed to align rRNA, several features have been added to allow alignment of conserved proteins (such as RecA, Hsp70 and ribosomal proteins) and construction of phylogenetic trees from protein alignments. Protein alignments are handled via ClustalW extensions that can be called from within the program, but such alignments are slower because the PT server cannot be used to speed up performance. Tree construction from protein alignments is handled by another extension, which calls some commands from the Phylogeny Inference Package (PHYLIP).

The workshop was extremely beneficial, both for teaching me the fundamentals of phylogeny and tree construction and for providing me with the technical expertise needed to use ARB. Thanks to the skills I acquired, I was able to install Linux and ARB, import and align 96 sequences of 16S rRNA genes from my metagenomic library using the Secondary Structure Editor, and use ARB-parsimony to insert these sequences into a tree and assess the diversity of microorganisms represented in the metagenomic library (Figure 1).

I am very grateful to the James Rennie Bequest for travel funding that enabled me to attend this workshop.

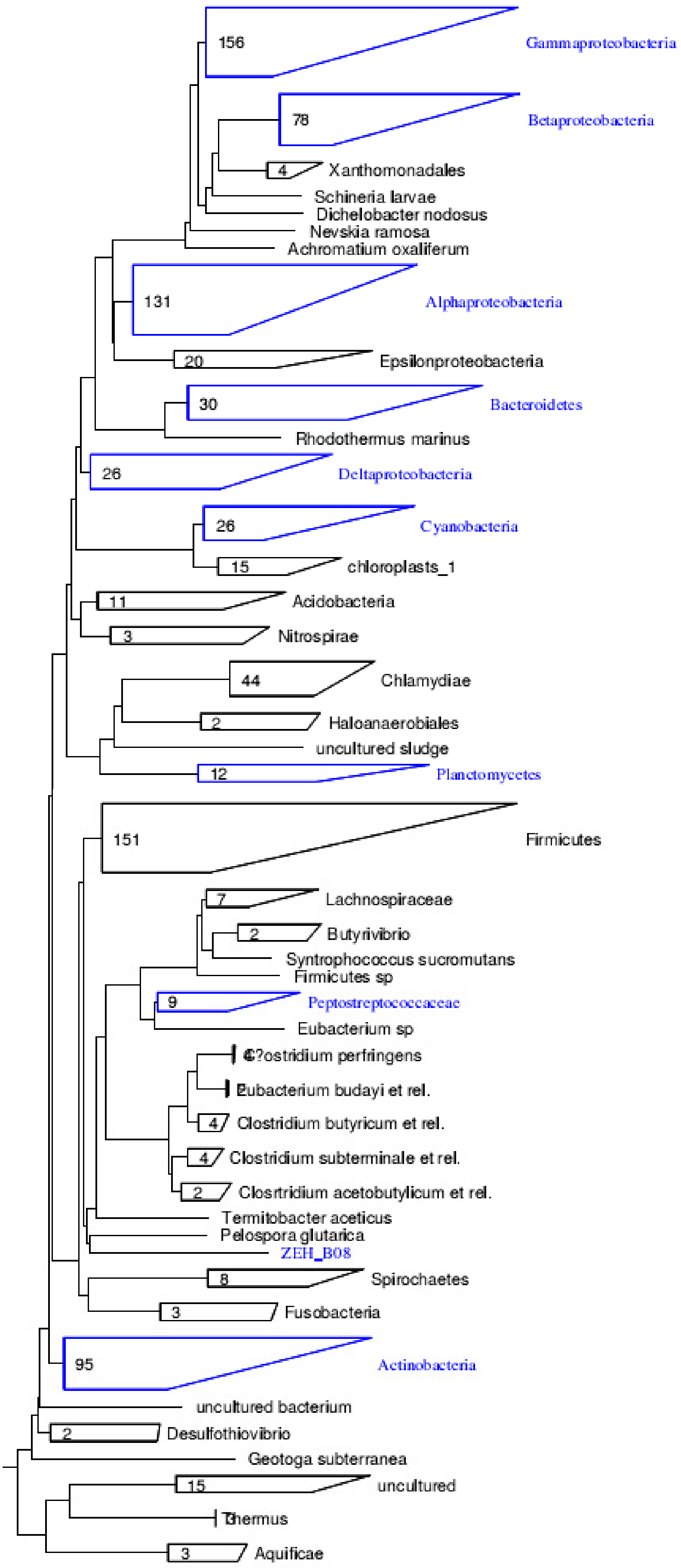

Figure 1. Phylogenetic tree showing the bacterial groups (in blue) within which microorganisms represented in the metagenomic library are distributed.