Development of new bioinformatic tools for finishing and annotating bacterial genomes: application to the genomic sequence of *P. profundum*: a barophile/ psycrophile bacterium

N. Vitulo, A. Vezzi, M. D'Angelo, A. Cestaro, P. Scannapieco, G. Valle

CRIBI, Università di Padova

The Genomic Research Group of CRIBI, University of Padova, is currently in the process of completing the genomic sequence of Photobacterium profundum, an extremophile bacterium adapted to life at high pressure and low temperature. The genome of this bacterium is more than 4 million base pairs long and about 99% of the sequence has now been obtained in our laboratory. We are currently in the finishing phase of the project, aiming to close the remaining 400 gaps (thus joining the current 400 contigs) and to confirm some genomic regions that are covered with low quality sequences.

Many computer programs were developed and implemented to carry out this project. The three main areas were related to project management, finishing and annotation process. In this presentation we cannot describe in details all these programs, therefore we will focus on two specific applications that were developed to help the finishing phase and in particular to identify adjacent contigs. Both methods are based on a comparative approach with other bacterial genomes.

The first approach is based on the assumption that at the finishing phase most gaps are relatively short and may be enclosed within a coding sequence. Since most of the coding sequences (about 70% in P. profundum) are encoding proteins that have homologs in other bacteria, we could search for contig ends that may be possibly coding for two regions of the same protein. With this premise we have developed a program that automatically performs a BlastX analysis on the extremity of all the contigs against a database of proteins obtained from the genomic sequences of more than 60 bacteria. The program performs a series of logical speculations to produce a series of suggestions on the possibility of joining together some contigs.

A second approach is based on the notion that several functionally related genes are often found next to each other in different genomes. Comparing the frequency that some genes are found in close proximity to other genes we can calculate the expectancy that some of these associations may also be present in P. profundum. In practical terms we have implemented a series of programs that perform the following tasks: 1) all the ORFs are identified; 2) BlastP of all the ORFs against the full set of proteins obtained from the genomic sequences 63 bacteria; 3) parsing the results and producing for each protein a 63-digit long binary string indicating the presence (1) or absence (0) of a homolog in the different bacteria; 4) storing the results of the above analysis in a SQL database; 5) systematic query of the database and identification of possible contigs that may be joined together.