A BIOINFORMATIC PIPELINE FOR GENE DETECTION AND FUNCTIONAL ANNOTATION OF PROTEIN SEQUENCES FROM METAGENOMIC SAMPLES VIA THE GALAXY PLATFORM

E. Ladoukakis\textsuperscript{1}, E. Pilalis\textsuperscript{2}, A. Chatziioannou\textsuperscript{2}, F. Kolisis\textsuperscript{1}

\textsuperscript{1}Laboratory of Biotechnology, School of Chemical Engineering, National Technical University of Athens, Greece
\textsuperscript{2}Institute of Biological Research and Biotechnology, National Hellenic Research Foundation, Athens, Greece

SUMMARY
In recent years the different DNA sequencing technologies have advanced greatly both in terms of reducing the time and cost per analysis as well as in achieving high yields (high-throughput techniques). Processing of the resulting data is a major challenge for bioinformatics due to its immense size (magnitude of 1-2GB per experiment), the dissimilarity between data formats from different technologies and the fact that several different bioinformatic tools need to be combined for each analytical task. In this work we have developed an automated workflow for analyzing metagenomic data by integrating numerous bioinformatic tools onto an open-source platform for biological and statistical analyses named Galaxy. This process comprises a) tools for genomic assembly from the metagenomic sequencing data, b) tools to identify putative open reading frames/genes in the examined data, c) homology-based tools for sequence annotation based on similarity searches against databases of curated protein sequences d) protein function prediction tools based on machine learning for the putative genes showing little or no homology to the already known ones. The integration of these tools onto the Galaxy platform was achieved by writing the appropriate scripts in Extensible Markup Language (xml) whereas the algorithms for handling input and output data exchange between tools, were developed in Perl and Python. Moreover, in order to facilitate the management of the massive amount of data, the Galaxy platform was connected with a MySQL database in such a way that the results can be available for user defined search queries as well as for export in a file format that is usable by a local installation of the above-mentioned database. This work is part of the research program "HotZyme" EU/FP7/KBBE-2010.3.5-04 on: "Systematic screening for novel hydrolases from hot environments" and aims to identify sequences of possible hydrolytic function of biotechnological interest from thermostable organisms.

INTRODUCTION
The field of genomics was established by Fred Sanger in the early 1950s with his successful sequencing of the B chain of insulin which was the first complete amino acid sequence ever to be determined\cite{1}. Since then, genomics earned its place as the cornerstone of molecular biology as it engulfs a wide range of disciplines, from wet lab techniques (e.g. DNA sequencing methods) to in silico methodologies in order to decipher the genomic content of any cells that can be cultured in vitro. A challenge to the vast potential of genomics appeared after the realization that most microorganisms cannot be grown readily in pure culture; in fact the percentage of an ambient microorganism population that can actually be cultured in standard conditions can differ from 0.1-1\%\cite{2, 3}.

The proposed solution for this predicament was DNA cloning directly from environmental samples and the construction of DNA libraries via the enrichment of the mixtures of organisms these samples encompass\cite{4}. Thus, a new approach emerged serving as the
evolution of genomics which became known by many names such as community genomics, environmental genomics, population genomics or by its more distinguished name, metagenomics[5]. Metagenomic analysis involves isolating DNA from an environmental sample, cloning the DNA into a suitable vector, transforming the clones into a host bacterium, and screening the resulting transformants. The clones can be screened for phylogenetic markers or “anchors,” such as 16S rRNA and recA, or for other conserved genes by hybridization or multiplex PCR[4] or for expression of specific traits, such as enzyme activity or antibiotic production[6-8] or they can be sequenced randomly. The subsequent exploitations of metagenomics include many industrial applications[9] (white biotechnology) where new biocatalysts can be discovered from terrestrial or marine environments as well environmental applications (green biotechnology) where comparative genomics can detect genetic signatures highly related with specific pollutants[10].

Unfortunately the numerous applications of metagenomics are somewhat thwarted by the complexity and cost of the implementation of each step of the analysis. The most popular DNA sequencing method which was introduced by Frank Sanger[11] (Sanger sequencing) and later adopted by the rest of the scientific community has been until recently a very expensive and time-consuming procedure despite its constant optimization. The field of metagenomics transitioned in a whole new era when high-throughput sequencing[12] techniques were introduced. These new techniques often referred to as next-generation sequencing (NGS)[13] offer greatly reduced costs (1000$ genome challenge) and operation times by means such as avoiding cloning, miniaturizing reactions, utilization of new chemical procedures, and exploitation of massively parallel sequencing.

These recent advances in sequencing technology have shifted the bottlenecks of a metagenomic analysis from the technical part onto the data management and annotation tasks. Next generation sequencing technologies due to their high-throughput nature tend to produce massive amounts of data (approximately 1-2GB per sequencing run) that need to be carefully processed and accurately interpreted in order to unveil their genomic content. In order to achieve a highly reliable annotation many different bioinformatic tools have been developed that handle each task of the analytical procedure; from assembling the raw sequencing data into larger nucleotide sequences[14] (contigs) and detecting putative open reading frames in them[15, 16], to identify similarities[17, 18] with other already known sequences in order to facilitate their function prediction. Nevertheless these tools usually require a very comprehensive knowledge of (bio)informatics and the necessary experience from the user in order to handle all the different types of data formats and to achieve an optimized analysis by entering the most appropriate input parameters each time. Furthermore the computational and disk space prerequisites of these tools tend to be far more demanding than the capabilities of a single personal computer. This creates the need for a computational infrastructure that can handle such computational tasks in an automated way and offer the graphic user interface through which the examining of the resulting data will be as intuitive as possible.

In this work we addressed this issue by developing an automated annotation workflow[19] for metagenomic samples, integrated into the Galaxy platform[20]. Galaxy (Figure 1.) is an open-source framework for the integration of computational tools and databases into a cohesive workspace and can be used for data intensive biomedical research. Galaxy provides a user-friendly web interface where users can develop, execute and share workflows of complex analyses that can be repeated on many different datasets, or refactored for different computing purposes. The users have the option of using a public server for their analysis or to install a
fully customizable local instance on their own server. The platform itself includes pre-configured tools for NGS analysis but also allows for integration of customized tools by each user with access to the server.

MATERIALS AND METHODS

The above-mentioned automated workflow (Figure 2.) when initialized, requires one or more sequencing datasets in FASTA or FASTQ format as well as the user defined specifications regarding the type of sequencing experiment (single or paired-end) and the file format. The rest of the input parameters for the tools in the workflow have been set up to their initial default values and require no user intervention. Nevertheless once the workflow is imported to a local Galaxy instance the user can access and fully customize it by modifying the parameters one by one in each module or even remove or add new modules in order to adjust it to the requirements of each new dataset. The tasks that are performed while the workflow is executed are the following:

Assembly: For the pipeline's assembler we chose Velvet[21] as it is one of the fastest and memory less demanding open source assemblers[22, 23]. Velvet can handle the raw sequencing data of the three leading sequencing technologies i.e. 454, Solexa and SOLID.

ORF detection: To detect open reading frames in the generated contigs as potential protein coding regions we used the Getorf program from the open source analysis package EMBoss[15].

Clustering: In a metagenomic sample originated from an environmental niche there is high probability that numerous conserved sequences will exist in the genomes of different species and consequently clustering is essential to avoid performing computationally intensive protein annotation analysis in identical or highly similar sequences. For this task we chose the CD-HIT[24] package with default clustering threshold of 0.9 i.e. 90% similarity.
Sequence homology analysis: The FASTA file with the representative sequences obtained by clustering is utilized as input to run a BLASTp[17] search against the UniProt[25] database. UniProt was chosen for its high quality protein annotations which can offer a starting point for the function prediction of the homologous sequences. The same file is also the input for a HMMER[18] search against the Pfam[26] database for domains of specific enzymatic activity to further assist in the curation of the sequences. The protein BLAST search provides homologous sequences at the protein level, which is the first step of functional annotation. Additionally HMMER algorithm uses Markov chain models[27] that render it very sensitive to remote homologies. Thus, it is a very efficient tool for the detection of protein functions even for sequences with very weak conservation compared to known proteins with established function.

EC number assignment: By exploiting the suite of tools provided by BioPerl[28], we constructed scripts that allowed us to create a list of all the annotated proteins in UniProt and their corresponding EC numbers. This list was imported in a MySQL[29] schema and was later called by another Perl script that we built and integrated into the Galaxy workflow that assigned putative EC numbers to sequences according to their BLAST hits from UniProt.

Orphans' separation: Considering the potential sequence novelty in a metagenomic sample we have build scripts using Biopython[30] and integrated them into the Galaxy workflow in order to parse the initial set of representative sequences and the BLAST result files, thus producing FASTA files with sequences of no homologies to known proteins. These sequences are then used again as BLASTp input but this time against the NCBI nr database in an attempt to find similarities even against sequences of lower quality annotation. The second BLAST results are also parsed returning a FASTA file of totally novel sequences.

Machine-learning protein function prediction: For the totally novel sequences showing no homology even against nr we use EFICAz[31] software that combines six different methods developed based on machine-learning practices in order to predict the putative protein function of the sequences in question. The resulting file contains the list of sequences followed by their putative EC numbers as predicted by the software's algorithms.

MySQL database: While each module of the pipeline is being executed the resulting datasets are imported in a MySQL schema, the dump of the corresponding tables is being generated and becomes available for download from the user. This is possible through the integration of the appropriate scripts that we developed in Python that parse the results, create the necessary tables in the MySQL database and populate the latter with the generated data of the modules that preceded them in the pipeline. Thus once the pipeline is terminated, the MySQL database is filled with a wide range of annotations related to putative functions of the metagenomics sequences.

In order for this pipeline to be developed we installed a local instance of Galaxy on NHRF's server (grissomdevweb.ekt.gr) as well as one on the Helios server of the University of Copenhagen both running on Debian 3.0.4. In addition we installed on the servers the necessary software (Velvet, EMBOSS, MySQL etc) and integrated the databases against which the homology searches would take place (nr, Pfam, UniProt). Both Galaxy instances were made remotely accessible through the appropriate configuration of the Apache web server of each distribution. The integration of the software tools that were exploited by each module of the pipeline was achieved by writing the appropriate scripts in Extensible Markup Language (xml)[32] providing thus, a friendly graphical user interface for each of them.
RESULTS AND DISCUSSION

The above-mentioned workflow was part of our bioinformatic contribution to the research program "HotZyme" which aims to identify sequences of possible hydrolytic function of biotechnological interest from thermostable organisms. For that purpose the pipeline was
further customized to include hydrolase databases and to narrow similarity searches against sequences of specific EC numbers according to their hydrolytic activity. For example, we developed the appropriate Perl scripts which allowed us to build a list of profile HMMs of the representative sequences of EC numbers starting with 3.-.-. (hydrolytic activity) from UniProt. Thus a new module was added to the pipeline that invoked HMMER search against a database of profile HMMs of hydrolases.

The first resulting data from the analysis of the project's metagenomic sample indicate a wide range of sequences of possible hydrolytic function as well as an even greater abundance in novel sequences that lacking any homology similarities to known proteins (data not shown).

Future additions to the pipeline include integration of additional databases against which similarity searches can take place (e.g. RefSeq), utilization of gene detection tools for prokaryotic and eukaryotic organisms, integration of tools for phylogenetic analysis and a front-end of the results' database which will provide the capability of indexing each sequence and associate it with the corresponding results of each tool. Furthermore as assembly tends to be a tedious and very demanding procedure that usually needs manual curation, we plan on removing the specific module from this workflow and construct a new workflow dedicated solely to the optimized utilization of different assemblers and the combination of their results to create a meta-assembly.

CONCLUSIONS
Metagenomics is an emerging field with potential for various industrial and environmental applications by exploiting the capabilities of high-throughput sequencing. The immense size of the generated data from these NGS technologies and the complexity of its annotation analysis have spawned the need for an automated bioinformatic application which will simplify data handling and will provide process optimization. The workflow presented here is a comprehensive tool for annotation of metagenomic datasets, starting from the raw output of DNA sequencers and leading to a complete database of functional annotation of the assembled sequences. The integration of the workflow into the Galaxy platform provides an intuitive and easily accessible tool for every biologist needing to handle very large datasets while having little or no expertise in (bio)informatics. Further development of the workflow will include additional tools for phylogenetic annotation, ready to apply advanced queries and availability of other genomic/protein databases as well as an integration of functional annotation tools that employ Gene Ontology and pathway enrichment analysis (e.g. StrAnGER[33]).

AKNOWLEDGEMENTS
This presented work has been funded by the project EU/FP7/KBBE-2010.3.5-04 "Systematic screening for novel hydrolases from hot environments" (HotZyme).

REFERENCES
9ο ΠΕΣΧΜ: Η Συμβολή της Χημικής Μηχανικής στην Αειφόρο Ανάπτυξη

Σχολή Χημικών Μηχανικών, Ε.Μ.Π., Αθήνα 23-25/5/2013