A Galaxy Workflow for the Functional Annotation of Metagenomic Samples

Eleftherios Pilalis[1], Eythymios Ladoukakis[2],
Fragiskos N. Kolisis[1,2], and Aristotelis Chatziioannou[1]

[1] Metabolic Engineering & Bioinformatics Group Institute of Biological Research and Biotechnology, National Hellenic Research Foundation, Athens, Greece
{epilalis,kolisis,achatzi}@eie.gr
[2] Laboratory of Biotechnology, School of Chemical Engineering, National Technical University of Athens, Athens, Greece

Abstract. In this work, an annotation workflow was developed, which performs a series of annotation tasks to sequences originating from metagenomic samples, using standard bioinformatics tools and Perl scripts. The Perl scripts interact with a Mysql database in order to store all annotation results to the respective tables, thus rendering easy the quick access and querying to all data. The whole pipeline was integrated into a Galaxy server, which provides a simple and intuitive interface that allows the user to easily create, run and share workflows for large datasets.

Keywords: Galaxy, Metagenomics, Gene Ontology Terms, Protein Function Prediction.

1 Introduction

Due to the recent advances in high-throughput sequencing, very large amounts of environmental DNA sequence data can be generated in a very short time. As a consequence, Metagenomics represent an emerging field with critical importance, as it describes the collection and analysis of the total DNA that is contained in an environmental niche [1]. Thus, sampling and analyzing environmental DNA is a promising way to identify novel genes, protein functions and enzymes, reflecting evolutionary adaptation to specific environmental conditions and therefore presenting great biotechnological and biomedical interest. However, the analysis of metagenomic data remains challenging, because of the complexity of microbial communities found in a particular environmental niche, the huge as well as heterogeneous genomic data volume and the inherent noise that characterize these DNA sequence data.

This annotation workflow is part of a large metagenomics 7th Framework project, named Hotzyme, aiming to investigate the global biodiversity in hot terrestial environments. Metagenomics has a great potential for assessing biodiversity and for enzyme discovery. This technology has been applied mainly to soil and marine water samples which revealed an enormous biological and molecular diversity. But to date,

I. Maglogiannis, V. Plagianakos, and I. Vlahavas (Eds.): SETN 2012, LNAI 7297, pp. 247–253, 2012.

very little work has been done on hot terrestrial environments, mainly due to the difficulty of access to various hot environments and the relatively lower concentration of biomass in such ecological systems [2]. This project aims to address this problem by screening for a new generation of (hyper)thermostable hydrolases from hot terrestial environments. Although thermostable hydrolases have been known for many years, the related research and applications have been limited to cultivated thermophilic microorganisms. Since most microorganisms (>99%) cannot easily be cultivated, many potentially active enzymes have never been characterized. This is particularly true for thermostable enzymes, since the number of isolated and characterized (hyper) thermophiles is very small. Therefore, the diversity of thermophiles and their encoded enzymes remains largely unexplored.

The analysis of metagenomic data begins with the assembly of the small DNA fragments, which are generated from the sequencer, to larger sequences called contigs. Then, the contigs are used as the main input to gene and protein function prediction algorithms. Although various tools have been developed for these purposes, it yet remains challenging from the computational aspect to handle efficiently the versatile different annotation tasks required. Therefore, we developed an automatic annotation workflow for metagenomic samples, integrated into the Galaxy platform. Galaxy [3] 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.

2 The Annotation Workflow

The pipeline (summarized in Figure 1) is initialized, complying with the specifications set by the user regarding the file containing the pair-end reads of the metagenomic experiment (fasta or fastq file). Subsequently, the following tasks are performed:

- Application of the Velvet program [4] for the assembly of the metagenomic reads to larger sequences (contigs)
- Parsing of the fasta file (contig_parser.pl) generated by the assembler, in order to populate the database with the basic information on the contigs (name, sequence and sample description included in the fasta file).

- Application of the Getorf program of the EMBOSS suite [5], in order to obtain putative open reading frames, within the contigs, that encode proteins. The output file of Getorf is parsed by a Perl script (getorf_parser.pl) and the results are stored into the Mysql database (table getorf).
- BlastN [6] search (nucleotidic blast) of the nucleotidic contig sequences against the Refseq [7] genomic database. The default threshold of high-scoring segment pairs (HSP) evalue is set to 0.001. The results of the Blast search are parsed by a Perl script (blastn_parser.pl) and are stored into the Mysql database (table blastn).
- Translation of the contigs to the corresponding aminoacids sequences (for all 6 reading frames) using the Transeq program of the EMBOSS suite, and submission of the peptidic sequences to BlastP [6] search (protein blast) against the Refseq protein database (default threshold HSP evalue set to 0.001). The results of the Blast search are parsed by a Perl script (blastp_parser.pl) and are stored into the Mysql database (table blastp).
- Submission of the peptidic sequences to a Hmmer [8] search against the Pfam [9] database (gathered threshold parameter), in order to obtain the annotation of the protein domains that are comprised in the peptides. The results of the Hmmer search are parsed by a Perl script (hmmer_parser.pl) and are stored into the Mysql database (table hmmer).
- Attribution of Gene Ontology terms [10] to the contigs, based on the protein homologs from the Refseq database (go_refseq.pl).

Consequently, once the pipeline is terminated, the Mysql database is filled with a wide range of annotations related to putative functions of the metagenomics sequences. The nucleotide blast search provides the information about the phylogenetic origin of the sequences that are found in the environmental sample. The protein blast search provides homologous sequences at the protein level, which is the first step of functional annotation. Additionally, the Hmmer search provides the independent protein domains that are found in the sample. It is noteworthy that the Hmmer algorithm uses Markov chain models 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. Finally, the database contains the Gene Ontology Terms that are attributed to the contigs on the basis of the detected homology with the proteins of the Refseq database. Homology is assessed using a maximum user-specified threshold of Blast HSP e-value. The Mysql database provides an easy and immediate access to all aforementioned results. All tables are linked by a unique contig id and hence all annotations of every contig can be easily and intuitively retrieved by SQL queries.

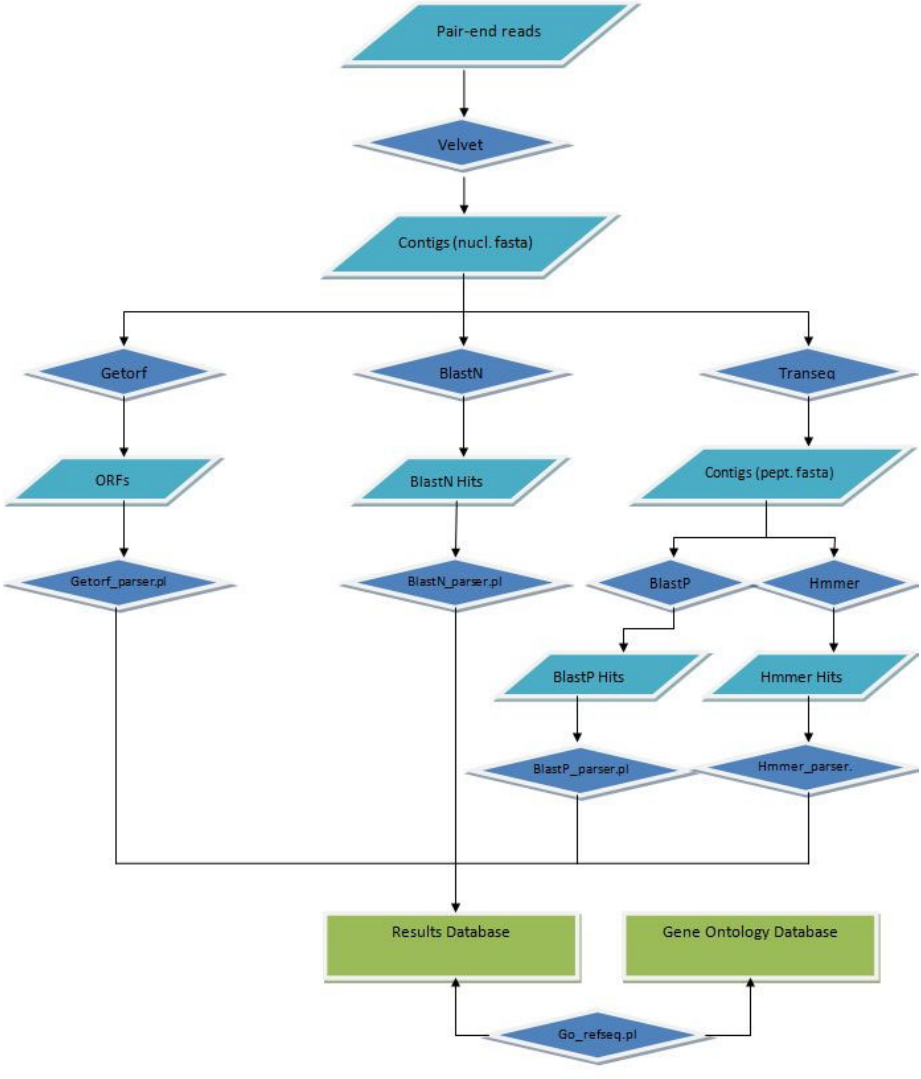

Fig. 1. The metagenomic annotation pipeline

3 Integration of Bioinformatics Tools/Algorithms

3.1 Stand-Alone Programs

- Velvet [4]: de novo metagenomic assembler based on de Bruijn graphs
- Blast + (Basic Local Alignment Search Tool) [6]: Finds regions of local similarity between sequences.

- Hmmer [8]: Implements profile hidden Markov models as probabilistic models to perform sequence comparisons against protein families (alignments) of the Pfam [9] database
- Emboss transeq [5]: Translates nucleotide sequences to peptide sequences
- Emboss getorf [5]: Finds and extracts open reading frames (ORFs) on input nucleotide sequences

3.2 Public Databases

- Refseq (genomic and protein) [7]: A comprehensive, non-redundant and well annotated database of genomes and proteins
- Pfam [9]: An extensive and highly curated database of protein domains, classified in families. It is a valuable resource of information on the functionality of query proteins

3.3 Scripts

contig_parser.pl

```
Usage: contig_parser.pl contig_file.fasta contig_file.log
db_name db_user_name db_password
```

- creates a Mysql database, the name of which is specified by the user (*db_name)*
- creates table *contig* into the database
- parses a fasta file containing the assembled contigs and fills table *contig* with all contigs

getorf_parser.pl

```
Usage: getorf_parser.pl getorf.out getorf_parser.log
db_name db_user_name db_password
```

- creates table *getorf* into the database
- parses a *Getorf* output file and stores the results into the *getorf* table

blastn_parser.pl

```
Usage: blastn_parser.pl blastn.out blastn_parser.log
db_name db_user_name db_password
```

- creates table *blastn* into the Mysql database
- parses a BlastN (nucleotice Blast) output file and stores the Blast results into the *blastn* table

blastp_parser.pl

```
Usage: blastp_parser.pl blastp.out blastp_parser.log
db_name db_user_name db_password
```

- creates table *blastp* into the Mysql database
- parses a BlastP (protein Blast) output file and stores the Blast results into the *blastp* table

hmmer_parser.pl

```
Usage: hmmer_parser.pl hmmer.out hmmer_parser.log db_name
db_user_name db_password
```

- creates table *hmmer* into the database
- parses a *Hmmer* output file and stores the results into the *hmmer* table

go_refseq.pl

```
Usage: go_refseq.pl hsp_evalue_threshold go_refseq.log
database user_name password
```

- creates table go_refseq into the database
- uses the BlastP results in order to retrieve the Gene Ontology Terms from the homologous proteins. Homology is assessed using a maximum user-specified threshold of blast HSP e-value (default 0.001), in the Refseq protein database. The Refseq annotations are contained in the Gene Ontology database, which was downloaded as a flat file from the Gene Ontology consortium website [11] and imported into a Mysql database called go.

4 Conclusion

Automated workflows are efficient tools for the performance of data mining tasks by users not expert in computing, but also for the efficient handling and analysis of very large datasets, such as metagenomic sequences. The workflow presented here is a comprehensive tool for annotation of metagenomic datasets, starting from the raw output of DNA sequencers (raw sequence reads) 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 with little to no expertise in bioinformatics. Further development of the workflow will include ready to apply advanced queries, availability of other genomic/protein databases an integration of functional annotation tools that employ Gene Ontology and pathway enrichment analysis (ex Stranger [12]).

Acknowledgements. The presented work in this paper has been funded by the EU/FP7/KBBE-2010.3.5-04 Microbial diversity and metagenomic mining for biotechnological innovation, "Systematic screening for novel hydrolases from hot environments" project (Hotzyme).

References

1. Desai, N., et al.: From genomics to metagenomics. Curr. Opin. Biotechnol. (2011)
2. Lorenz, P., Eck, J.: Metagenomics and industrial applications. Nat. Rev. Microbiol. 3(6), 510–516 (2005)
3. Giardine, B., et al.: Galaxy: a platform for interactive large-scale genome analysis. Genome Res. 15(10), 1451–1455 (2005)

4. Zerbino, D.R., Birney, E.: Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18(5), 821–829 (2008)
5. Rice, P., Longden, I., Bleasby, A.: EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 16(6), 276–277 (2000)
6. Altschul, S.F., et al.: Basic local alignment search tool. J. Mol. Biol. 215(3), 403–410 (1990)
7. Pruitt, K.D., Tatusova, T., Maglott, D.R.: NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 35(Database issue), D61–D65 (2007)
8. Finn, R.D., Clements, J., Eddy, S.R.: HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 39(Web Server issue), W29–W37 (2011)
9. Bateman, A., et al.: The Pfam protein families database. Nucleic Acids Res. 32(Database issue), D138–D141 (2004)
10. Ashburner, M., et al.: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25(1), 25–29 (2000)
11. `http://www.geneontology.org/GO.downloads.database.shtml`
12. Chatziioannou, A.A., Moulos, P.: Exploiting Statistical Methodologies and Controlled Vocabularies for Prioritized Functional Analysis of Genomic Experiments: the StRAnGER Web Application. Front Neurosci. 5, 8 (2011)