Marine Microbial Biodiversity, Bioinformatics Biotechnology

Grant agreement n°287589

Acronym : Micro B3
Start date of project: 01/01/2012, funded for 48 months

Deliverable D5.8

Report describing processing pipelines and associated software

Version:3
Circulated to: Authors and Coordinator (23.12.13)
Approved by:
Expected Submission Date: 24.12.2013
Actual submission Date: 24.12.2013

Lead Party for Deliverable: Guy Cochrane, EMBL-EBI

Mail: **cochrane@ebi.ac.uk** **Tel.: +44 1223 492564**

Dissemination level:	
Public (PU)	X
Restricted to other programme participants (including the Commission Services) (PP)	
Restricted to a group specified by the consortium (including the Commission Services) (RE)	
Confidential, only for members of the consortium (including the Commission Services) (CO)	

Generalist summary

Detailed analysis of the sequences derived from marine samples and isolated marine organisms is required to turn high volumes of data into useful, scientifically meaningful information, ready to be interpreted by the marine science community. Four Micro-B3 partner institutions have built, and now operate, expert analysis pipelines, covering metagenomes, prokaryotes, eukaryotes and viruses, providing information on genes, functions and taxonomic diversity. Sequence data from two of the Micro-B3 flagship projects, Ocean Sampling Day and Tara Oceans, are being analysed through these pipelines. In this deliverable report, we describe these pipelines with a view to helping marine scientists to evaluate the information the data can provide and to further discussions that will lead to accuracy and efficiency improvements in the pipelines.

Contents

Generalist summary 2
Outline 3
Aims and future work 3
Implementations 4
The pipelines 5
Metagenomics 5
SILVAngs 9
Protists 13
Viruses 16
Prokaryotes 23

The Micro B3 project is funded from the European Union's Seventh Framework Programme (Joint Call OCEAN.2011-2: Marine microbial diversity – new insights into marine ecosystems functioning and its biotechnological potential) under the grant agreement no 287589. The Micro B3 project is solely responsible for this publication. It does not represent the opinion of the EU. The EU is not responsible for any use that might be made of data appearing herein.

Outline

In this report, we describe analytical pipelines in operation under Micro-B3. These pipelines consume nucleic acid sequence data and provide, for isolated organisms, genome-scale functional annotation and, for anonymous environmental samples, assessments of both functional potential and taxonomic diversity. Covering prokaryotes, protists and viruses, the pipelines represent the state-of-the-art in marine molecular analyses and serve within the Micro-B3 project as core feeds into the Micro-B3 Information System (see figure I). This report details the analytical steps in each of the pipelines, making some reference to the design process that led to the inclusion of the steps.

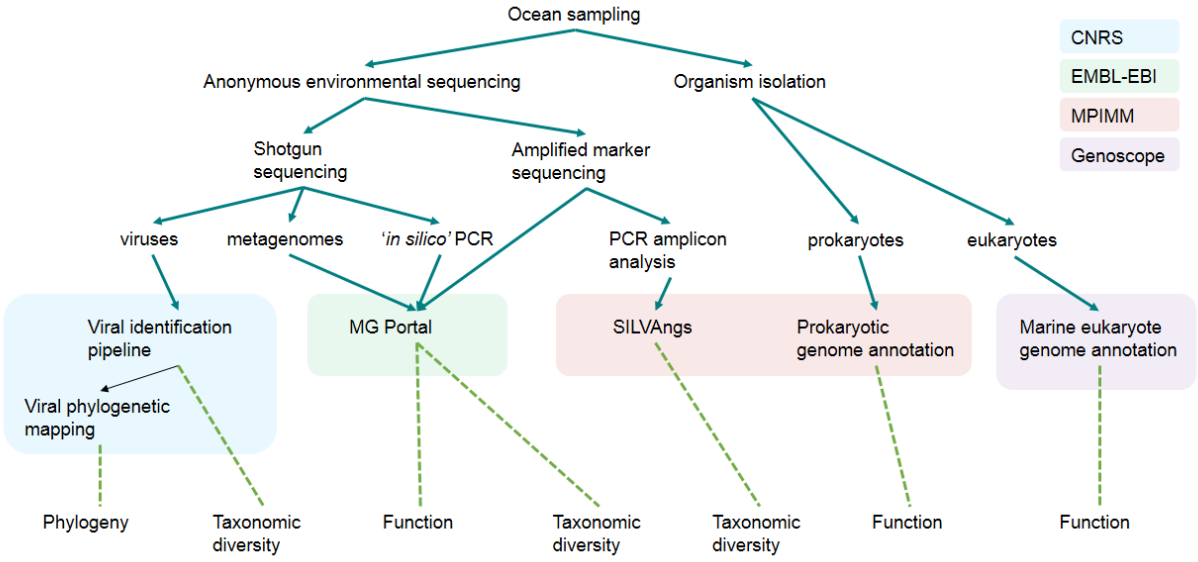

FIGURE I: MICRO-B3 ANALYSIS OPTIONS, SHOWING ROUTING THROUGH D5.8 CONTRIBUTOR ANALYTICAL PIPELINES TO INFORMATION OUTPUTS

Aims and future work

In this report, our aim is to provide useful descriptions of functional and taxonomic analytical pipelines, each of which is a series of processing and analytical steps that allow the inference of biologically meaningful information from input data. While this report does not specifically cover contextual data (for which the reader should refer to Micro-B3 workpackage 4 for details of standards and protocols), the authors stress their importance for any useful interpretation of the information outputs of the analytical pipelines.
We intend a number of uses for this report.

- We wish to describe to Micro-B3 users (such as Ocean Sampling Day sampling groups) the processes through which their data flow, such that they are able to consume and assess outputs with appropriate confidence.
- In laying out the steps, we aim to encourage a technical discussion between the designers of the analytical pipelines relating to the choice and tuning of the steps and driving at greater accuracy, performance and utility.
- We intend to provide a perspective on analytical approaches for the consumption of those involved in software development and the provision of services beyond Micro-B3, with a view to greater interoperability between methods and data intermediates.

Implementations

The analytical pipelines described in this document support Micro-B3 uses, including OSD and Tara Oceans. The EMBL-EBI metagenomic functional and taxonomic analyses (EMBL-EBI Metagenomics) and the MPIMM taxonomic analysis[1] (SILVAngs) are provided as open public services at https://www.ebi.ac.uk/metagenomics/ and https://www.arb-silva.de/ngs/, respectively. These services are appropriate in particular for OSD methods, where a broad diversity of users will approach analysis directly, and we expect individually, at dispersed times. The Genoscope protist annotation pipeline, which has its dominant Micro-B3 use in the Tara Oceans project, provides an internal face to incoming sequence data from the institute's sequencing machines. The CNRS viral analysis pipeline faces usage specifically from this same data source and captures specific data directly from Genoscope.

[1] The taxonomic analysis pipelines provided by the EMBL-EBI MG Portal (EMBL-EBI Metagenomics) and the SILVAngs sytstem are conceptually very different; specifically, the EMBL-EBI system, based upon QIIME, takes a clustering approach and in post-analysis provides mappings into known taxonomic classifications, while the MPIMM SILVAngs system takes a more phylogenetic approach, connecting sequences with known taxa independently of cluster relationships.

The pipelines

Metagenomics

Summary of functionality

The pipeline accepts raw nucleotide reads as input. It then performs quality control; feature (rRNA and protein) prediction; function prediction and taxonomic prediction, as detailed below. The full analysis results are made available for download and visualisation via the EBI metagenomics web interface (www.ebi.ac.uk/metagenomics). An overview of the pipeline is given in Figure II.

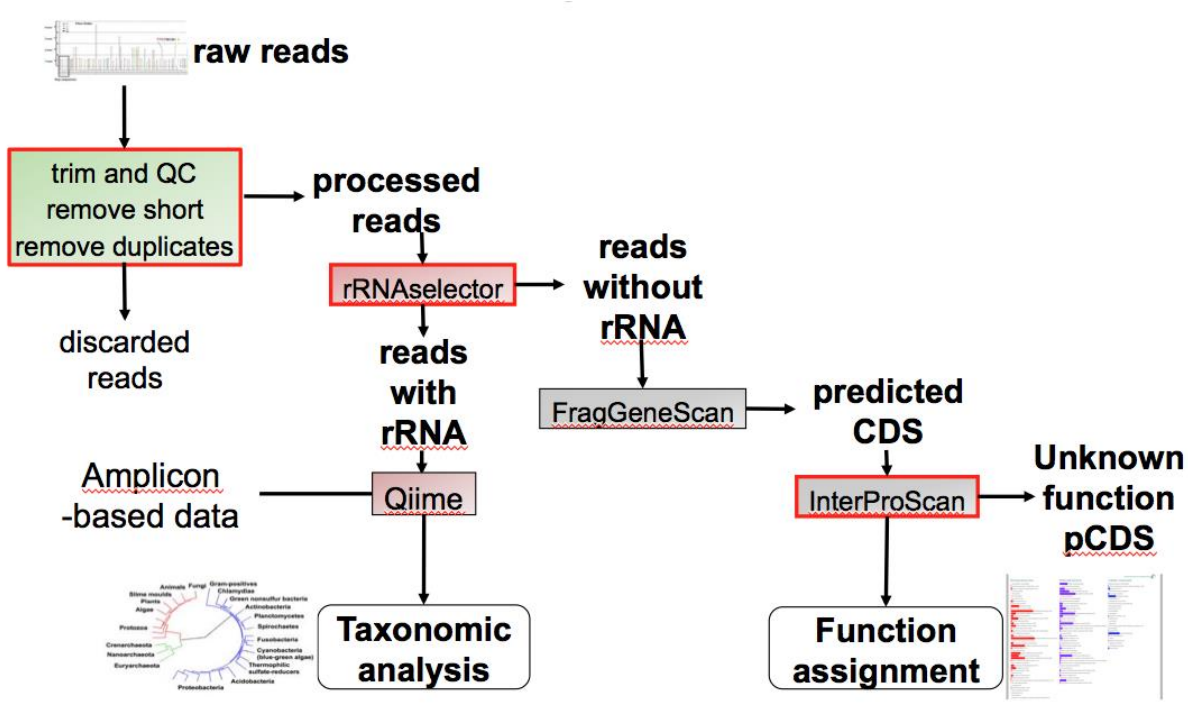

FIGURE II: AN OVERVIEW OF THE METAGENOMICS PIPELINE

Technical description

Input: The pipeline takes as input raw nucleotide reads from the following formats and platforms:

- FASTQ format (all platforms)
- SFF format (454 and Ion Torrent),
- SOLiD csfasta/qual format
- PacBio HDF5 format

In addition, FASTA files can be submitted, although as they do not contain quality information, not all of the quality control steps described below can be performed.

Illumina paired-end reads are merged together using SeqPrep (http://github.com/jstjohn/SeqPrep) prior to being submitted to the pipeline, but otherwise, reads are not assembled. Quality control is not required at the submission stage, since this is performed as part of the pipeline.

Pipeline processing: The pipeline is fully automated and has a Python framework that manages the execution of individual steps on a computer cluster via IBM's Platform Load Sharing Facility (LSF) and collates results files together in a simple directory structure. In its

current format, the pipeline is optimised to the EBI LSF configuration and therefore not readily adaptable to other platforms. However, the pipeline is currently being ported to run on the Taverna Workflow Management System (Wolstencroft *et al.*, 2013), in order to make it more modular, flexible and render the source code easy to share with the wider metagenomics community.

The pipeline runs the following steps:

1. Quality Control (QC): Different QC steps are performed, depending on the sequencing platform used to generate the nucleotide reads. The QC steps are intended to remove low-quality and uninformative reads from the data set, so that they are not needlessly passed on to the later stages of the pipeline. The QC steps include: read trimming (to remove adapters, etc., using BioPython sff-trim (Cock *et al.*, 2009) and trimmomatic (Lohse *et al.*, 2012)), removal of ambiguous leading/trailing bases, removal of reads shorter than 100 nucleotides, removal of reads where the proportion of ambiguous bases >=10%, clustering to remove duplicate sequences (using UCLUST v1.1.579 (Edgar, 2010) or using pick_otus.py from QIIME v1.5 (Caporaso, 2010) for prefix-based filtering) and repeat masking using RepeatMasker open-3.2.2 (http://www.repeatmasker.org).
2. Feature prediction: The pipeline predicts both ribosomal RNA coding (using models from rRNAselector v1.0.0 (Lee *et al.*, 2011)) and protein coding (using FragGeneScan v1.15 (Rho *et al.*, 2010)) features. Predicted rRNAs are used for taxonomic analysis, particularly in the absence of marker gene studies for the same sample, and predicted protein coding sequences (pCDS) are fed into the functional analysis steps in the pipeline.
3. Function prediction: Function prediction is performed by analysing the predicted coding sequences using InterProScan 5 (Hunter *et al.*, 2012). Not all of InterPro's member databases are designed to work on fragmentary data or microbial sequences, therefore only a subset of the 11 available InterPro database analyses are run. These are: CATH-Gene3D v3.5.0 (Lees *et al.*, 2012); PRINTS v41.1 (Attwood *et al.*, 2012); Pfam v24.0 (Punta *et al.*, 2012); TIGRFAMs v9.0 (Haft *et al.*, 2013) and PROSITE Patterns v20.66 (Sigrist *et al.*, 2012). Gene Ontology (GO) terms (The Gene Ontology Consortium, 2000) for molecular function, biological process and cellular component are associated to the pCDSs by virtue of the InterPro2GO mapping service (Burge *et al.*, 2012).
4. Taxonomic diversity prediction: QIIME v1.5 is used to classify reads into Operational Taxonomic Units (OTUs), in order to give an indication of the diversity of species found in a particular sample. Subsequently, the RDP classifier (Cole *et al.*, 2009) and the Greengenes reference database (DeSantis *et al.*, 2006) are used for classification of archaeal and bacterial species.

Output: The complete set of information about a metagenome, including the raw, submitted sequence data, metadata and analysis results is made freely available for download and visualisation via the EBI metagenomics web interface.

Interoperability

The pipeline provides output in a range of formats, so that it can be connected with other elements within the MicroB3 project. Formats include FASTA for sequence files, tab-separated values (TSV) for functional results, Biom and TSV formats for OTU tables, and Newick format for phylogenetic trees. Functional results in GFF3 format will also be made available.

Metagenomics pipeline references

Wolstencroft,K., Haines,R., and Fellows,D. (2013) The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. Nucleic Acids Res., 41, W557–W561 doi:10.1093/nar/gkt328

Cock,P.J., Antao,T., Chang,J.T., Chapman,B.A., Cox,C.J., Dalke,A., Friedberg,I., Hamelryck,T., Kauff,F., Wilczynski,B. and de Hoon,M.J. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics, 25, 1422-1423 doi:10.1093/bioinformatics/btp163

Lohse,M., Bolger,A.M., Nagel,A., Fernie,A.R., Lunn,J.E., Stitt,M. and Usadel,B. (2012) RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. Nucleic Acids Res., 40, W622-627 doi:10.1093/nar/gks540

Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics, 26, 2460-2461 doi:10.1093/bioinformatics/btq461

Caporaso,J.G., Kuczynski,J., Stombaugh,J., Bittinger,K., Bushman,F.D., Costello,E.K., Fierer,N., Peña,A.G., Goodrich,J.K., Gordon,J.I. et al (2010) QIIME allows analysis of high-throughput community sequencing data. Nat. Methods, 7, 335-336 doi:10.1038/nmeth.f.303

Lee,J.H., Yi,H. and Chun,J. (2011) rRNASelector: a computer program for selecting ribosomal RNA encoding sequences from metagenomic and metatranscriptomic shotgun libraries. J. Microbiol., 49, 689-691 doi:10.1007/s12275-011-1213-z

Rho,M., Tang,H. and Ye,Y. (2010) FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Res., 38, e191 doi:10.1093/nar/gkq747

Hunter,S., Jones,P., Mitchell,A.L, Apweiler,R., Attwood,T.K., Bateman,A.G., Bernard,T., Binns,D., Bork,P., Burge,S., et al. (2012). InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res., 40, D306–D312 doi:10.1093/nar/gkr948

Lees,J., Yeats,C., Perkins,J., Sillitoe,I., Rentzsch,R., Dessailly,B.H. and Orengo,C. (2012) Gene3D: a domain-based resource for comparative genomics, functional annotation and protein network analysis. Nucleic Acids Res., 40, D465-D471 doi:10.1093/nar/gkr1181

Attwood,T.K., Coletta,A., Muirhead,G., Pavlopoulou,A., Philippou,P.B., Popov,I., Romá-Mateo,C., Theodosiou,A. and Mitchell,A.L. (2012) The PRINTS database: a fine-grained protein sequence annotation and analysis resource--its status in 2012. Database (Oxford). 2012:bas019 doi:10.1093/database/bas019

Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang N., Forslund, K., Ceric, G., Clements, J. et al. (2012) The Pfam protein families database. Nucleic Acids Res. 40, D290-D301 doi:10.1093/nar/gkr1065

Haft,D.H., Selengut,J.D., Richter,R.A., Harkins,D., Basu,M.K. and Beck,E. (2013) TIGRFAMs and Genome Properties in 2013. Nucleic Acids Res., 41, D387-D395 doi:10.1093/nar/gks1234

Sigrist,C.J.A., de Castro,E., Cerutti,L., Cuche,B.A., Hulo,N., Bridge,A., Bougueleret,L. and Xenarios,I. (2012) New and continuing developments at PROSITE. Nucleic Acids Res., 41, D344-D347 doi:10.1093/nar/gks1067

The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. Nat. Genetics, 25, 25-29

Burge,S., Kelly,E., Lonsdale,D., Mutowo-Muellenet,P., McAnulla,C., Mitchell,A.L., Sangrador-Vegas,A., Yong,S., Mulder,N. and Hunter,S. (2012) Manual GO annotation of predictive protein signatures: the InterPro approach to GO curation. Database (Oxford), 2012 doi:10.1093/database/bar068

Cole,J.R., Wang,Q., Cardenas,E., Fish,J., Chai,B., Farris,R.J., Kulam-Syed-Mohideen,A.S., McGarrell,D.M., Marsh,T., Garrity,G.M. and Tiedje,J.M. (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res., 37, D141-D145 doi:10.1093/nar/gkn879

DeSantis,T.Z., Hugenholtz,P., Larsen,N., Rojas,M., Brodie,E.L., Keller,K., Huber,T., Dalevi,D., Hu,P. and Andersen,G. L. (2006). Greengenes, a chimera-checked 16S rRNA Gene database and workbench compatible with ARB. Appl. Environ. Microbiol., 72, 5069-5072 doi:10.1128/AEM.03006-05

SILVAngs

Summary of functionality
SILVAngs (SILVA Next Generation Sequencing) is a new service to meet user requests to provide a fast and accurate classification service for next generation sequencing data. SILVAngs accepts any kind of short and long read sequence data in FASTA format, performs quality control, alignment and taxonomic classification of ribosomal RNA gene sequences based on the curated SILVA taxonomy. All steps (upload, progress monitoring, visualisation of results and download of data) can be steered by the SILVAngs web interface. The system has access control with the possibility to share project data between users. Further developments cover the entry of contextual (meta) data fully compliant with the minimum information standards of the Genomic Standards Consortium and the direct submission of results to ENA for data sharing and long-term storage. The system is accessible at www.arb-silva.de/ngs.

Technical description
The pipeline can be divided into the following steps (see Figure III):

- Alignment
- Quality management
- De-replication (identification of identical sequences)
- Clustering at a user-defined threshold (OTU definition)
- Classification of the OTUs/reads

SILVAngs reflects the common process of analysing rDNA amplicon reads from next generation sequencing platforms. However, there are some specific features of SILVAngs compared to similar software pipelines in terms of quality and efficiency.

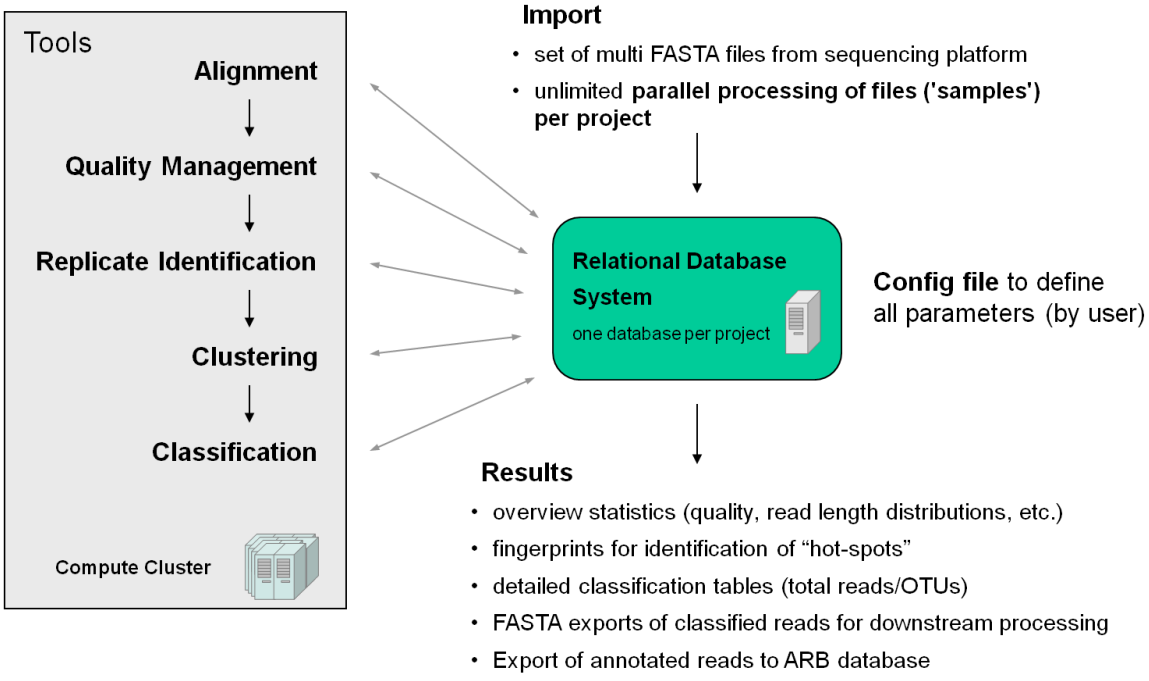

FIGURE III: SILVANGS ARCHTIECTURAL OVERVIEW, INCLUDING DEPICTION OF MAIN PIPELINE STEPS (LEFT GREY BOX).

Input: The pipeline accepts input data in multi-fasta format in which each input file represents one sample. Several samples that belong to a project (e.g. a transect, time series, etc.) should be organised and uploaded to the same SILVAngs project. The name of a sample is defined by the file name of the multi-fasta input file. Files in Standard Flowgram Format (.sff), e.g. from 454 sequencing, and quality values, e.g. in FASTQ files, are not accepted. Resolution of short rDNA reads is limited anyway and adequate quality control is assured by the subsequent steps of the analysis process.
Alignment: In the first step all input reads are aligned by the SILVA aligner (http://www.arb-silva.de/aligner/sina-download/ (Pruesse *et al.*, 2012)). A number of measures are taken by the aligner; problematic reads (such as PCR artefacts) or even contamination of the data set with non-rDNA sequences are identified and corresponding reads filtered out. The numbers of rejected reads during initial alignment (separated into classes) are given in the final statistics of the analysis pipeline.
The alignment step allows exports of aligned sequences for all reads which entered the classification step of the pipeline. These can be used for a detailed inspection e.g. with the ARB software package (Ludwig *et al.*, 2004) or other sequence editors.
Further quality management: All reads which have not been rejected by the previous alignment step undergo further quality filtering including length, ambiguity and homopolymer checks. These are the common parameters which allow quality control on the level of the primary sequence information. The length cut-off can be defined by the user, whereas for ambiguities and homopolymers the same thresholds as for the SILVA databases are used (max. 2%). Notably, only the aligned region of each read is considered and further processed. Therefore, non-rDNA overhangs do not bias the results.
All reads with insufficient quality values are not considered for further processing – again, the number of reads rejected during quality management is indicated in the final statistics of the analysis.
De-replication (identification of identical sequences): All remaining reads enter the de-replication stage of the pipeline. 100% identical reads, ignoring overhangs, are identified and only the longest read is further processed.
This is a common approach to reduce calculation time. The exclusion of identical reads provides information on the microdiversity within a dataset beneath the level of taxonomic assignment - in case of OTU definition, no information could be obtained on how “clonal” the input dataset or selected taxonomic units/groups are.
Clustering/OTU definition: Technically, this is just another de-replication step to further reduce the number of reads that needs to be classified. Compared to previous de-replication, clustering is done on a 97-99% identity level which can be adjusted by the user. This is motivated by the fact that sequencing errors and operon heterogeneities can introduce 3% artificial divergence in the data. This step gives OTU numbers. It even allows comparison of OTUs across different samples.
Classification of OTUs/reads: In the classification step the representative reads are compared to the SILVA reference datasets of the small (16S/18S) and large (23S/28S) subunit rDNA with its corresponding SILVA taxonomy (Quast *et al.*, 2013). Currently, and unlike other pipelines, the SILVAngs pipeline uses a BLAST-based approach for classifying the reads according to the SILVA taxonomy.
Since the SILVA reference datasets are comprehensive, quality-controlled, and of high integrity, and the corresponding taxonomy is phylogeny-based and manually curated, SILVAngs fully relies on best BLAST hit and avoids more complex procedures such as least common ancestor (LCA) approaches. Only significant hits (see below) are considered;

everything else is assigned to a class called 'No Relative' and is offered to the user for further inspection. However, usually just a few percent of the reads within a complete dataset go into this class, since the majority of reads is normally classified.
The classification result of each representative read (including the additional class 'No Relative') is finally mapped back to all other members (reads) of the OTU cluster and to the corresponding identical reads from the de-replication step.
SILVAngs assumes that SILVA's phylogeny-based taxonomy has a reliable resolution down to the genus level. We urge caution in interpretation of species assignments, which need to be supported by careful inspection of the corresponding reads. The reliability of classification is assured by two features of the system:

- Avoidance of using solely the best BLAST hit; this can be far away from the query sequence (e.g. >5% divergence) or the reported identity can be artificial because of partial BLAST alignments; to circumvent these shortcomings, best hits are tested and only accepted if (sequence identity + alignment coverage) / 2 >= 93, a threshold which has been determined empirically and provides an acceptable compromise between sensitivity and accuracy.
- SILVAngs does not refer to any non-standardized user-based annotation (from original submission of the reference sequence). Only the SILVA taxonomic information is taken to identify the unknown OTU/read. In other words, all accepted reads of the project are finally mapped to the standardized SILVA taxonomy. This is the actual result of the pipeline visualized by various views and supported by additional statistics and exports.

Interoperability
SILVAngs produces several outputs for visualization, reporting and documentation. However, the single most important output type for Micro B3-IS and Ocean Sampling day are the 'fingerprint' files (see Figure IV). If this file gets converted to a classical abundance data table (also commonly named sites-by-species table), it will be interoperable with the EATME tool of WP5, general R scripts, and will function as the main input to the BioVeL workflows for the Ocean Sampling Day (http://www.microb3.eu/news/new-axis-collaboration-biovel-workflows-micro-b3-ocean-sampling-day).

	A	B	C	D	E	F	G
1	1	2	3	4	5		
2	0	0	0	1	0	Bacteria;Acidobacteria;Acidobacteria;Subgroup 6;	
3	0	0	1	0	0	Bacteria;Actinobacteria;Acidimicrobiia;Acidimicrobiales;OCS155 marine group;	
4	0	0	0	0	1	Bacteria;Actinobacteria;Acidimicrobiia;Acidimicrobiales;Sva0996 marine group;	
5	0	0	5	0	0	Bacteria;Actinobacteria;Actinobacteria;Corynebacteriales;Corynebacteriaceae;Corynebacterium;	
6	0	0	3	0	0	Bacteria;Actinobacteria;Actinobacteria;Corynebacteriales;uncultured;	
7	0	1	0	0	0	Bacteria;Actinobacteria;Actinobacteria;Micrococcales;Microbacteriaceae;Leucobacter;	
8	0	1	0	0	0	Bacteria;Actinobacteria;Actinobacteria;Micrococcales;Micrococcaceae;Micrococcus;	
9	0	0	9	0	0	Bacteria;Actinobacteria;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Propionibacterium;	
10	0	0	0	1	0	Bacteria;Actinobacteria;Rubrobacteria;Rubrobacterales;Rubrobacteriaceae;Rubrobacter;	
11	0	0	0	2	0	Bacteria;Actinobacteria;Thermoleophilia;Gaiellales;uncultured;	
12	0	0	0	2	0	Bacteria;Bacteroidetes;Cytophagia;Cytophagales;Cytophagaceae;Flexibacter;	
13	0	0	0	32	2	Bacteria;Bacteroidetes;Cytophagia;Order II Incertae Sedis;Rhodothermaceae;Rubricoccus;	
14	0	0	0	3	0	Bacteria;Bacteroidetes;Cytophagia;Order II Incertae Sedis;Rhodothermaceae;uncultured;	
15	0	0	0	0	2	Bacteria;Bacteroidetes;Cytophagia;Order III Incertae Sedis;BIgi5;	
16	0	0	1	0	0	Bacteria;Bacteroidetes;Flavobacteria;Flavobacteriales;Cryomorphaceae;Fluviicola;	
17	0	0	0	10	0	Bacteria;Bacteroidetes;Flavobacteria;Flavobacteriales;Cryomorphaceae;NS7 marine group;	
18	4	0	0	12	54	Bacteria;Bacteroidetes;Flavobacteria;Flavobacteriales;Cryomorphaceae;Owenweeksia;	
19	0	0	0	0	10	Bacteria;Bacteroidetes;Flavobacteria;Flavobacteriales;Flavobacteriaceae;Aequorivita;	
20	0	0	0	1	0	Bacteria;Bacteroidetes;Flavobacteria;Flavobacteriales;Flavobacteriaceae;Algibacter;	
21	0	0	0	1	0	Bacteria;Bacteroidetes;Flavobacteria;Flavobacteriales;Flavobacteriaceae;Arenibacter;	
22	0	0	1	0	0	Bacteria;Bacteroidetes;Flavobacteria;Flavobacteriales;Flavobacteriaceae;Cloacibacterium;	
23	0	0	0	10	15	Bacteria;Bacteroidetes;Flavobacteria;Flavobacteriales;Flavobacteriaceae;Flavobacterium;	
24	0	0	0	70	0	Bacteria;Bacteroidetes;Flavobacteria;Flavobacteriales;Flavobacteriaceae;Gaetbulibacter;	
25	0	0	0	1	0	Bacteria;Bacteroidetes;Flavobacteria;Flavobacteriales;Flavobacteriaceae;Gelidibacter;	
26	0	0	0	2	0	Bacteria;Bacteroidetes;Flavobacteria;Flavobacteriales;Flavobacteriaceae;Lutibacter;	
27	0	0	1	0	0	Bacteria;Bacteroidetes;Flavobacteria;Flavobacteriales;Flavobacteriaceae;Maritimimonas;	

FIGURE IV: EXAMPLE 'FINGERPRINT' OUTPUT FILE OF SILVANGS. THE HEADER ROW DEPICTS THE SAMPLE NAMES (HERE SIMPLY NAMED 1-5). EACH OTHER ROW CONTAINS THE COUNT OF A SPECIFIC TAXON IN A SAMPLE. THE NAME OF THE TAXON IS GIVEN IN THE LAST COLUMN.

SILVAngs pipeline references

Christian Quast, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research, 41(D1):D590{D596, 2013. doi: 10.1093/ nar/gks1219.

Anna Klindworth, Elmar Pruesse, Timmy Schweer, Jörg Peplies, Christian Quast, Matthias Horn, and Frank Oliver Glöckner. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Research, 41(1):e1, 2013. doi: 10.1093/nar/gks808.

Elmar Pruesse, Jörg Peplies, and Frank Oliver Glöckner. SINA: accurate high throughput multiple sequence alignment of ribosomal rna genes. Bioinformatics, 2012. doi: 10.1093/bioinformatics/bts252.

Protists

Context
Marine eukaryotes are weakly represented in public genomic databases. This is especially true for planktonic species that are targeted in current projects such as Tara Oceans. The paucity of knowledge on these organisms challenges gene prediction since standard methods rely on training on descriptions of known genes. Many parameters describing the exon/intron structure, lengths and number of exons and introns, frequencies of splicing sites, and even the genetic code can differ between species.
Public availability of transcriptome and metatranscriptome sequence data is increasing. This might represent an excellent opportunity to detect coding sequences, and splicing information without any *a priori* knowledge. However, this increase will also exacerbate CPU intensiveness and impact the computational feasibility of the pipeline.

Objective
Our aim is to develop a procedure to detect genes from a given eukaryote genome taking in account the following constraints:

- Existence of minimal (or no) *a priori* information
- Maximum automation
- Optimization of CPU time, adaptable on each query depending on associate resources
- Output of a set of genes in a standard format
- Production of a final gene model that integrates number and types of evidence that are variable according to genomes

Rationale
The pipeline is made of a suite of independent tools producing different gene prediction evidence in order to maximize the probability of detection of each gene. An overview is given on figure V. Since these methods provide alternative outputs, a central stage is a reconciliation procedure. The final gene prediction is an exon/intron structure defined by a list of coordinates of start and ends of exons. The pipeline is also able to predict UTR regions.

Technical description
cDNA alignments: The alignment process is multi-step. The number of modern transcriptomic reads combined with their relatively short length (about 100nt) impacts accuracy and computational time. So, transcriptomic reads are assembled in contigs using a dedicated assembler. At this stage we use Oases (Schulz *et al.*, 2012). Subsequently, contigs are aligned and mapped on the reference genomic sequence using blat (Kent, 2002). At this step, we apply some filters. In case of abundant resources we keep contigs that are aligned on more than 80% of their length with at least 95% of identity. To get more precise splice site definitions, we realign each contig on the same region (we locate the relevant genomic region based on the previous blat alignment) with est2genome (Mott, 2012).
An alternative to this strategy consists of direct alignment of cDNA reads onto genomic sequence using Star software (Dobin *et al.*, 1997) and prediction of full transcript models with Gmorse5.

Protein alignments: We have defined a protein database containing Uniprot records and proteins from complete marine genomes that have not yet been integrated into Uniprot.

Comparisons and alignments are processed here in two steps. A first direct and rapid comparison is done using blat (Kent, 2002). Then, a second alignment is performed using GeneWise (Birney *et al.*, 2004). Since blat is quite stringent and because some genomes are phylogeneticaly very distant from Uniprot records, we run a supplementary round of comparison between proteins and genomic regions lacking any alignment with blat. In this round we use classical blast instead of blat.

***ab initio* gene prediction**: Even if resources in protein and cDNA are abundant, it is important to include an *ab initio* prediction for rarely expressed genes. This tool is even more important on genomes from unexplored phyla, those that are phylogeneticaly distant from the any known to Uniprot and those for which no transcriptomic data are available. We use snap (Korf, 2004) which is based on a Markov model and requires calibration on representative exon/intron structures. For this calibration, we use information from spliced alignments obtained with cDNA or proteins.

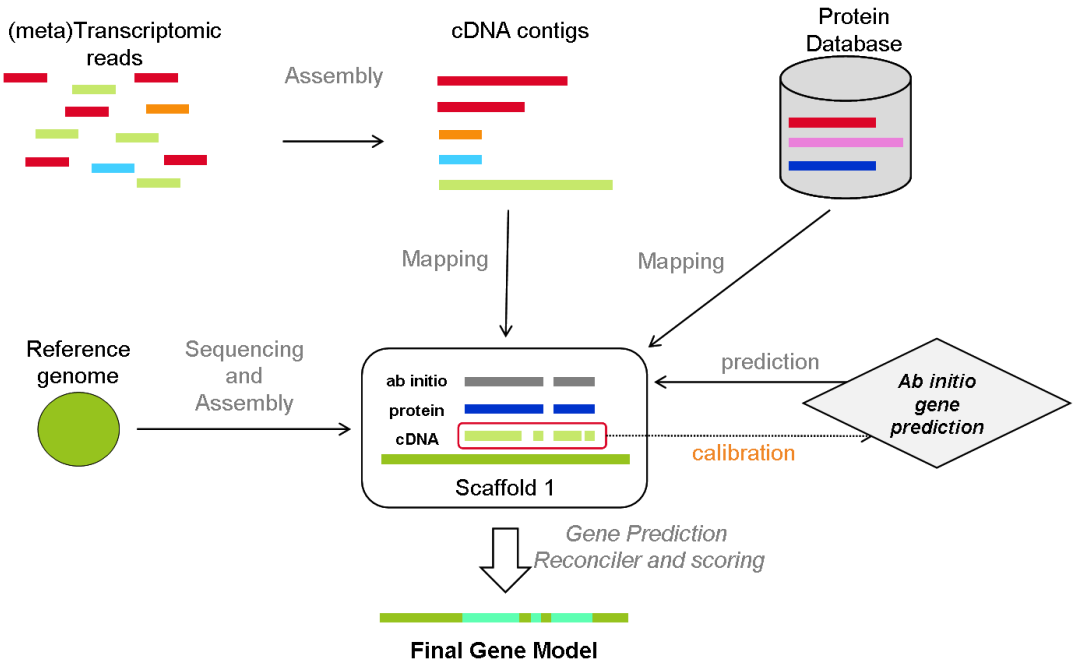

FIGURE V: GENE PREDICTION BY RECONCILIATION OF DIFFERENT EVIDENCE SOURCES.

Reconciliation: The final reconciliation of the above mentioned methods uses gaze (Howe, 2002). This tool is a framework that allows a complete flexibility which is well adapted for eukaryote genomes showing various gene structures. Genoscope has a long experience with this tool, having use it for the annotation of genomes from various lineages (tunicate, teleost fish, ciliate, fungi, stramenopile, plant, etc.). The framework provides a score on each gene model, representing the sum of individual scores of the sources, weighted according to a defined. Final output of gene predictions is stored in GFF3 format. A direct visualization can be made using a genome browser (www.gbrowse.org).

Comments

Overall, the global procedure is largely inspired by that used at Genoscope over years to annotate numerous eukaryotes. Methods are wrapped in scripts for automation and benefit from the local IT architecture. We are continuously improving the technical specification, in terms of computation and automation. One caveat in using Gaze is that it requires precise and sensitive configuration, for which specific experience is needed. A further limitation is

the lack of support for alternative splicing. However, in the context of a first description of an unknown lineage, we accept this issue. We are, though, exploring technical alternatives to compensate for these limitations. In parallel, technical optimization must be applied to reduce the total CPU time, which ranges from 15 to 30 CPU days per genome. One approach would be to reduce protein and cDNA database size without losing information.
The pipeline as described here does not give any functional annotation on genes. To overcome this, we will add a layer integrating tools that predict functional domains such as Interproscan (Zbodnov and Apweiler, 2001)) or CD-Search (Marchler-Bauer and Bryant, 2004; Fong and Marchler-Bauer, 2008). Analysis and detection of orthology and paralogy relationships between genes can also be added in future.

Protists pipeline references

Schulz, M. H., Zerbino, D. R., Vingron, M. & Birney, E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinforma. Oxf. Engl. 28, 1086–1092 (2012).

Kent, W. J. BLAT--the BLAST-like alignment tool. Genome Res. 12, 656–664 (2002).

Mott, R. EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. Comput. Appl. Biosci. CABIOS 13, 477–478 (1997).

Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. Bioinforma. Oxf. Engl. 29, 15–21 (2013).

Denoeud, F. et al. Annotating genomes with massive-scale RNA sequencing. Genome Biol. 9, R175 (2008).

Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. Genome Res. 14, 988–995 (2004).

Korf, I. Gene finding in novel genomes. BMC Bioinformatics 5, 59 (2004).

Howe, K. L., Chothia, T. & Durbin, R. GAZE: a generic framework for the integration of gene-prediction data by dynamic programming. Genome Res. 12, 1418–1427 (2002).

Zdobnov, E. M. & Apweiler, R. InterProScan--an integration platform for the signature-recognition methods in InterPro. Bioinforma. Oxf. Engl. 17, 847–848 (2001).

Marchler-Bauer, A. & Bryant, S. H. CD-Search: protein domain annotations on the fly. Nucleic Acids Res. 32, W327–331 (2004).

Fong, J. H. & Marchler-Bauer, A. Protein subfamily assignment using the Conserved Domain Database. BMC Res. Notes 1, 114 (2008).

Viruses

As efficient killers of plankton, viruses are recognized to have a significant ecological role in marine ecosystems (Suttle, 2007; Breitbart, 2012). Nevertheless, due to their extreme diversity and variability and lack of universal marker genes, marine viruses outside of phages are still poorly catalogued. The first virus pipeline module developed by IGS (CNRS partner 6) is designed to identify sequences of likely viral origin in environmental shotgun ('metagenomic') sequence data.

Amongst viroplankton, the Nucleo-Cytoplasmic Large DNA Viruses (NCLDVs) have been shown to be abundant in marine environments (Monier *et al.*, 2008; Hingamp *et al*., 2013). NCLDVs constitute an apparently monophyletic group that consists of at least 6 families of viruses: *Poxviridae*, *Asfarviridae*, *Iridoviridae*, *Ascoviridae*, *Phycodnaviridae*, and *Megaviridae* (Yutin and Koonin, 2012; Arslan *et al.*, 2011). IGS developed a second virus pipeline module which, based on phylogenetic mapping of NCLDV marker genes, provides tools for estimating NCLDV family-level diversity in environmental shotgun data.

Together, these two modules constitute the virus annotation pipeline component of D5.8.

Summary of functionality

The MB3-Virus pipeline is made of two modules that may be used independently, or chained together as a complete pipeline (Figure VI). The first module, referred to as the "Virus Identification Pipeline" assigns taxonomic annotations to environmental reads or contigs (pre-assembled reads) of putative viral origin. The method underlying the Virus Identification Pipeline is a sequence similarity-based taxonomy assignation method developed by IGS specifically for accurate identification of viruses. It is somewhat similar to the method applied by MEGAN (Huson *et al.*, 2007) but uses an adaptive E-value threshold specific for each protein.

The second module, referred to as "NCLDVs Phylogenetic Mapping" phylogenetically maps NCLDVs pre-assigned reads or contigs on a set of 24 NCLDV marker genes. Phylogenetic mapping (Monier *et al.*, 2008) is a method used to place and classify a new sequence (usually a short environmental sequence) within a reference tree using a precompiled multiple sequence alignment. So far only 2 multiple sequence alignment sets (type B DNA polymerase: PolB and DNA mismatch repair protein: MutS) have been used to assign metagenomic reads to NCLDVs (Hingamp *et al*., 2013). Here IGS provides a complete protocol for phylogenetic mapping of metagenomes, as well as a set of reference alignments for 24 NCLDV marker genes.

The scientific validity of both modules of the virus pipeline was tested as part of the WP6 D6.2 - annotation of at least two metagenomics datasets – one of which (Tara Oceans pyrosequence dataset) was peer-reviewed and published (Hingamp *et al*., 2013).

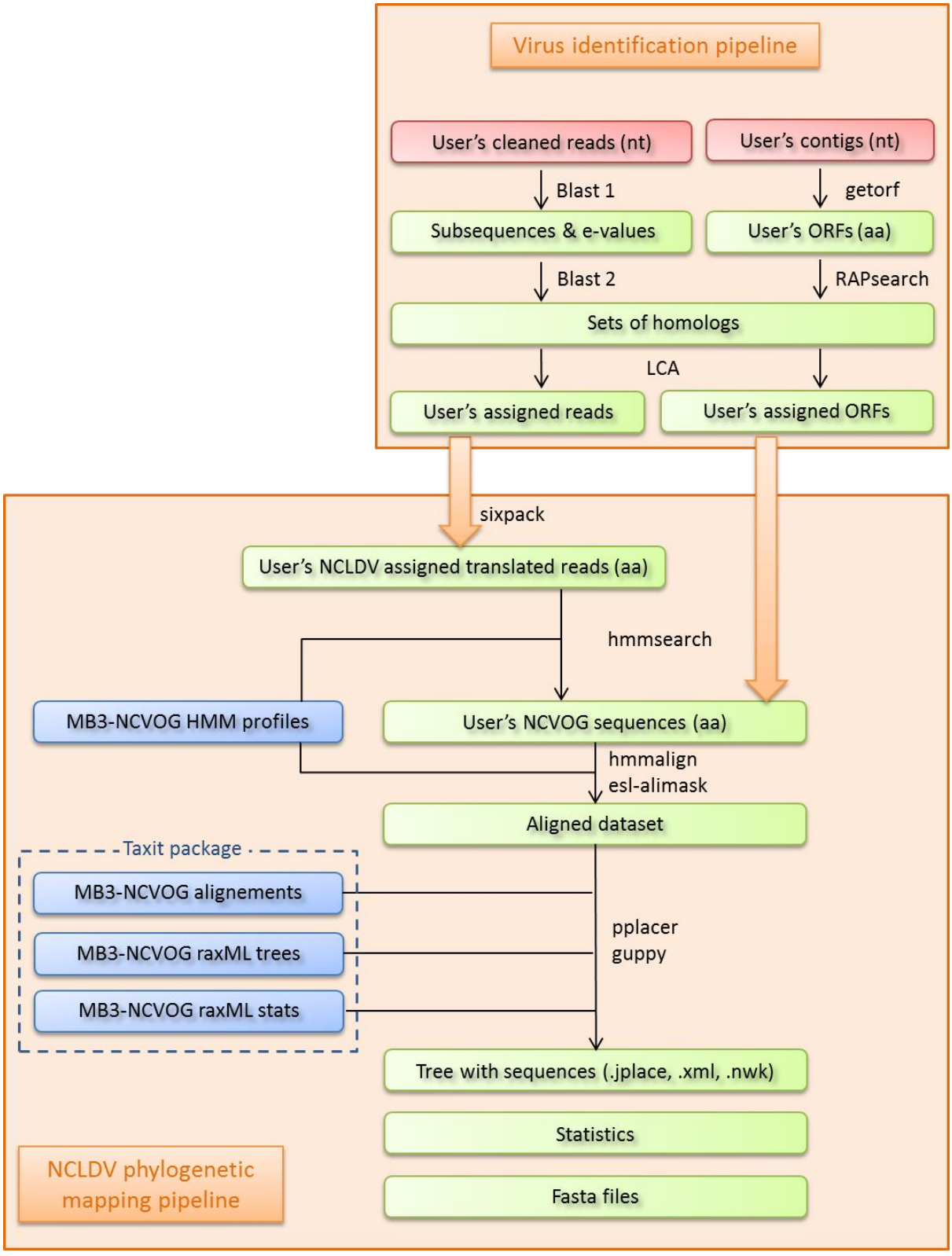

FIGURE VI: THE MB3-VIRUS PIPELINE. USER INPUT FILES ARE IN RED, OUTPUT FILES IN GREEN, AND MB3 PROVIDED REFERENCE DATASETS IN BLUE.

Technical description of Virus Identification Pipeline

Input: The input data consists of:

- Metagenome shotgun sequences: should be either cleaned reads (for 2bLCA protocol below) or contigs (1rLCA protocol below) in FASTA format.

- Reference database: The user may use any reference database in FASTA format.The most recent version of the Uniref100 non redundant protein database is recommended as Last Common Ancestor (LCA) taxa are provided for each protein cluster (ftp://ftp.ebi.ac.uk/pub/databases/uniprot/uniref/uniref100/).
- Taxonomic database providing the taxonomic information of the genes of the reference database (ftp://ftp.ncbi.nih.gov/pub/taxonomy/).

Pipeline processing: Whatever the form of the metagenome sequences (reads or contigs), the taxonomic annotation can be carried out according to the initial 2bLCA method as published in Hingamp *et al.*, 2013. For contig annotations, IGS proposes the more recent 1rLCA method, as yet unpublished, but which provides comparable accuracy with a significant reduction in computation time.

2bLCA protocol: This taxonomy annotation method follows 3 steps:

- A first BLASTx (B1) against UniRef100 to retrieve the corresponding B1 E-value (at least < 10-5), as well as the subsequence of the UniRef100 subject fragment aligned in the best scoring segment pair.
- A second BLASTx (B2) of the UniRef100 subject subsequence aligned in B1 against the same UniRef100. All sequences with a E-value < to the B1 E-value are retained to constitute a set of homologs for the read.
- A reduction to the Last Common Ancestor of the homolog set is then performed to assign taxonomic annotation to the read.

This 2bLCA protocol, well suited for the accurate annotation of small to medium scale metagenome datasets (Hingamp *et al.*, 2013), becomes computationally expensive for large datasets because of the required second BLAST.

1rLCA protocol: An alternative method was investigated which would scale in the case of recent large datasets, such as produced by Illumina HiSEQ2000 sequencers. This annotation process follows two steps:

- A first comparison of the query protein sequences to a reference database, typically UniRef. In the IGS implementation, Rapsearch2 was used, a Blast-like software, which is up to 1000 times faster than Blast (Ye *et al.*, 2011). For each unknown protein, a log-evalue threshold is set, which by default is equal to 90% of the log-E-value of the best hit for that protein (there are many options available for the user to set this threshold). Hits with a log-E-value lower than the threshold are considered as potential homologs of the unknown protein.
- In a second step, taxonomic annotations are computed for each query protein, based on the LCA of its potential homologs (the set of best Rapsearch hits).

A benchmark study shows that for protein sequences, 1rLCA is both a lot faster and more reliable than 2bLCA.

Output: The output is an annotation file with 4 columns (query gene ID, taxonomic assignment, log10 evalue of the best hit, number of hits used to compute the LCA).

Interoperability: The output can be used to reduce a full metagenome data set to the subset of sequences assigned to NCLDVs. This filtered subset is then suitable for submission to the NCLDV phylogenetic mapping second module in order to produce fine-grained taxonomic assignations. It is worth noting that this first module output can also be used directly as a source of coarse-grained taxonomic information for plankton diversity studies.

Technical description of Virus Identification Pipeline

Input: The input should be FASTA files of amino-acid sequences. These can be amino-acid sequences derived from 6 frame translation of NCLDV assigned reads, or Open Reading Frames (ORF) of contigs.

Reference dataset provided MB3-NCVOG Taxit Package: The main part of the virus pipeline construction consisted of building reference trees for NCLDV marker genes. Once built as described below, these resources are used as references onto which phylogenetic mapping can be carried out as described under "Pipeline Processing" below.

1. Retrieving NCLDV genomes: IGS searched in GenBank complete genome sequences of viruses of a size between 300 and 3000 kb, or viruses belonging to *Poxviridae*, *Ascoviridae*, *Mimiviridae*, *Asfarviridae*, *Iridoviridae* or *Phycodnaviridae*. We obtained 330 accession numbers, and discarded redundant sequences (same TaxID, different strains) to obtain a set of 119 NCLDV genomes (see table 1).

2. Retrieving NCVOG in the new NCLDV genome sequence set: IGS selected 28 Nucleo-Cytoplasmic Virus Orthologous Groups (NCVOG) used in previous study on NCLDVs (Hingamp *et al.*, 2013; Yutin and Koonin, 2013): 17 being conserved in most NCLDVS, 11 being conserved in *Megaviridae* and *Phycodnaviridae*. We performed a blastx of NCLDV genome sequences on previous published NCVOG sequences (Yutin *et al.*, 2009), available at ftp://ftp.ncbi.nih.gov/pub/wolf/COGs/NCVOG/. When hit positions correspond to annotated CDS, CDS sequences were retrieved. As all genomes were not completely annotated, IGS also searched HMM profiles of the Yutin's NCVOG on intergenic ORFs. ORFs were retrieved using getorf from EMBOSS suite (see D6.1), HMM were built and searched with the HMMER software (Finn *et al.*, 2011).

Intergenic ORFs having a significant match (e-value < 10^{-3}) with a HMM profile were also retrieved. Finally, very few blastx hits were not validated (Table 2). Non-validated hits were frequently associated with low blastx scores.

Family	Nb of genomes
Ascoviridae	5
Asfarviridae	5
Iridoviridae	25
Marseilleviridae	3
Megaviridae	10
Pandoraviridae	2
Phycodnaviridae	17
Poxviridae	52

TABLE 1: NUMBERS OF NCLDV COMPLETE GENOMES USED BY FAMILY.

	Validated by CDS hits	Validated by HMM	Not Validated
NCVOG0022	112	5	1
NCVOG0023	116	6	1
NCVOG0038	113	6	1
NCVOG0040	76	6	0
NCVOG0052	104	6	1
NCVOG0076	101	4	0
NCVOG0158	53	2	0
NCVOG0236	118	4	1
NCVOG0249	113	5	1
NCVOG0262	113	5	1
NCVOG0271	98	5	1
NCVOG0272	101	9	1
NCVOG0274	98	7	1
NCVOG0276	99	5	1
NCVOG0278	98	4	2
NCVOG0313	26	1	0
NCVOG0330	82	8	1
NCVOG1068	80	4	0
NCVOG1115	61	6	0
NCVOG1127	32	0	1
NCVOG1129	24	0	0
NCVOG1137	25	1	3
NCVOG1164	109	4	1
NCVOG1166	57	5	1
NCVOG1192	38	2	5
NCVOG1216	24	0	0
NCVOG1342	26	0	4
NCVOG1353	78	0	2
NCVOG1354	55	1	1

TABLE 2: NUMBERS OF SEQUENCES RETRIEVED AFTER VALIDATION BY MATCH ON CDS OR HMM PROFILES AND NUMBERS OF SEQUENCES DISCARDED.

3. New NCVOG alignments and reference tree building: Sequences were aligned using T-Coffee (Notredame *et al.*, 2000), using default parameters. Alignments were manually curated by IGS, discarding sequences with gaps on the conserved sites of other aligned sequences. Four alignments displaying very few conserved sites were discarded.
Phylogenetic trees were built using RaxML (Stamatakis, 2006), using default parameters.
HMM profiles were built with the HMMER software (hmmbuild routine, Finn *et al.*, 2011).
Files provided for each of the 24 remaining NCVOG are grouped in a Taxit package (http://fhcrc.github.io/taxtastic/):

- Reference alignment: NCVOGXXXX.aln
- RaxML tree: RaxML_result.NCVOGXXXX
- RaxML info: RaxML_info.NCVOGXXXX
- Sequences informations: NCVOGXXXX.csv

HMM profiles (NCVOGXXXX.hmm) are provided separately.

Pipeline processing: First, tools from the HMMER suite are used to construct the aligned dataset:

1. hmmsearch retrieves input amino-acid sequences matching with MB3-NCVOG HMM profiles. Here the e-value threshold is adjustable, by default $<10^{-3}$.
2. hmmalign aligns sequences on the HMM profile.
3. esl-alimask reformats the alignment: all columns that are not gaps in the reference annotation will be kept and all other columns will be removed.
1. Then sequences are placed on the reference tree:
4. Pplacer uses the reference alignment provided, RaxML trees and infos files provided to compute the placement of the target sequence on the reference tree.
5. Guppy generates output trees
6. Perl scripts generate statistic tables and groups sequences by placed nodes in fasta files.

A test version from step 2 to 5 is already available on a *beta* version of the website phylogeny.fr (http://www.phylogeny.fr/phylo_cgi_svn/index.cgi). The overall pipeline is written in bash, and some intermediate scripts in Perl. There are no settings adjustable by the user. The program versions are the following:

HMMER 3.0 (March 2010)

pplacer v1.1.alpha11rc1-0-g52f06eb

guppy v1.1.alpha11rc1-0-g52f06eb

Output: Trees are provided in .jplace, newick and xml formats.

Tabular text files provide:

- Summary of the number of reads placed on each node.
- The list of query sequences placed on nodes and their associated probabilities.

Sequences are sorted and split by node in fasta format (tar-gzipped).

Interoperability: These outputs are directly useful for interpretation and figure drawing.

Virus pipeline references

Arslan, D., Legendre, M., Seltzer, V., Abergel, C., and Claverie, J.-M. (2011). Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. PNAS 108, 17486–17491.

Finn, R.D., Clements, J., and Eddy, S.R. (2011). HMMER web server: interactive sequence similarity searching. Nucl. Acids Res. 39, W29–W37.

Hingamp, P., Grimsley, N., Acinas, S.G., Clerissi, C., Subirana, L., Poulain, J., Ferrera, I., Sarmento, H., Villar, E., Lima-Mendez, G., et al. (2013). Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. ISME J 7, 1678–1695.

Huson, D.H., Auch, A.F., Qi, J., and Schuster, S.C. (2007). MEGAN analysis of metagenomic data. Genome Res. 17, 377–386.

Koonin, E.V., and Yutin, N. (2010). Origin and Evolution of Eukaryotic Large Nucleo-Cytoplasmic DNA Viruses. Intervirology 53, 284–292.

Monier, A., Claverie, J.-M., and Ogata, H. (2008). Taxonomic distribution of large DNA viruses in the sea. Genome Biology 9, R106.

Notredame, C., Higgins, D.G., and Heringa, J. (2000). T-coffee: a novel method for fast and accurate multiple sequence alignment. Journal of Molecular Biology 302, 205–217.

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22, 2688–2690.

Ye, Y., Choi, J.-H., and Tang, H. (2011) RAPSearch: a Fast Protein Similarity Search Tool for Short Reads. BMC Bioinformatics 12:159.

Yutin, N., and Koonin, E.V. (2013). Pandoraviruses are highly derived phycodnaviruses. Biology Direct 8, 25.

Yutin, N., Wolf, Y.I., Raoult, D., and Koonin, E.V. (2009). Eukaryotic large nucleo-cytoplasmic DNA viruses: Clusters of orthologous genes and reconstruction of viral genome evolution. Virology Journal 6, 223.

Prokaryotes

Summary of functionality
The MB3-Prokaryotic pipeline is an in-house hosted pipeline for the analysis and automatic annotation of assembled bacterial and archaeal genomes. Several analysis steps predict Open Reading Frames (ORF), transmembrane helices or signal peptides. Based on sequence similarity searches, the predicted ORFs are annotated functionally. The results of the structural and functional annotation are returned to the user in GFF3 format. Additionally, the results of the annotation pipeline can be further investigated using the JCoast software.

Technical description
As a first step, a FASTA file containing the genome sequence is uploaded and gene prediction is carried out using the Glimmer3 software (Delcher *et al*., 2007). All subsequent steps are based on GenDB version 2.2.1 (Meyer, 2003): A new GenDB project is created and the predicted ORFs are loaded into GenDB. For each predicted ORF, BLAST similarity searches are performed against several sequence databases including NCBI-nr (Altschul, 1997), Swiss-Prot (Boeckmann, 2003), KEGG (Kanehisa, 2002), COG 66 (Tatusov, 2000) and genomesDB (Richter *et al*., 2008).
Similarity searches for protein families are done using databases from Pfam 27.0 (Finn *et al.*, 2010) and InterPro (Hunter *et al.*, 2011).
SignalP (Emanuelsson et al. 2007) is used for signal peptide prediction and TMHMM (Krogh *et al.*, 2001) for transmembrane helix-analysis. The MicHanThi (Quast, 2007) software predicts gene functions based on the results of the similarity searches against NCBI-nr (including Swiss-Prot) and InterPro. The output format of the annotation results is GFF3.

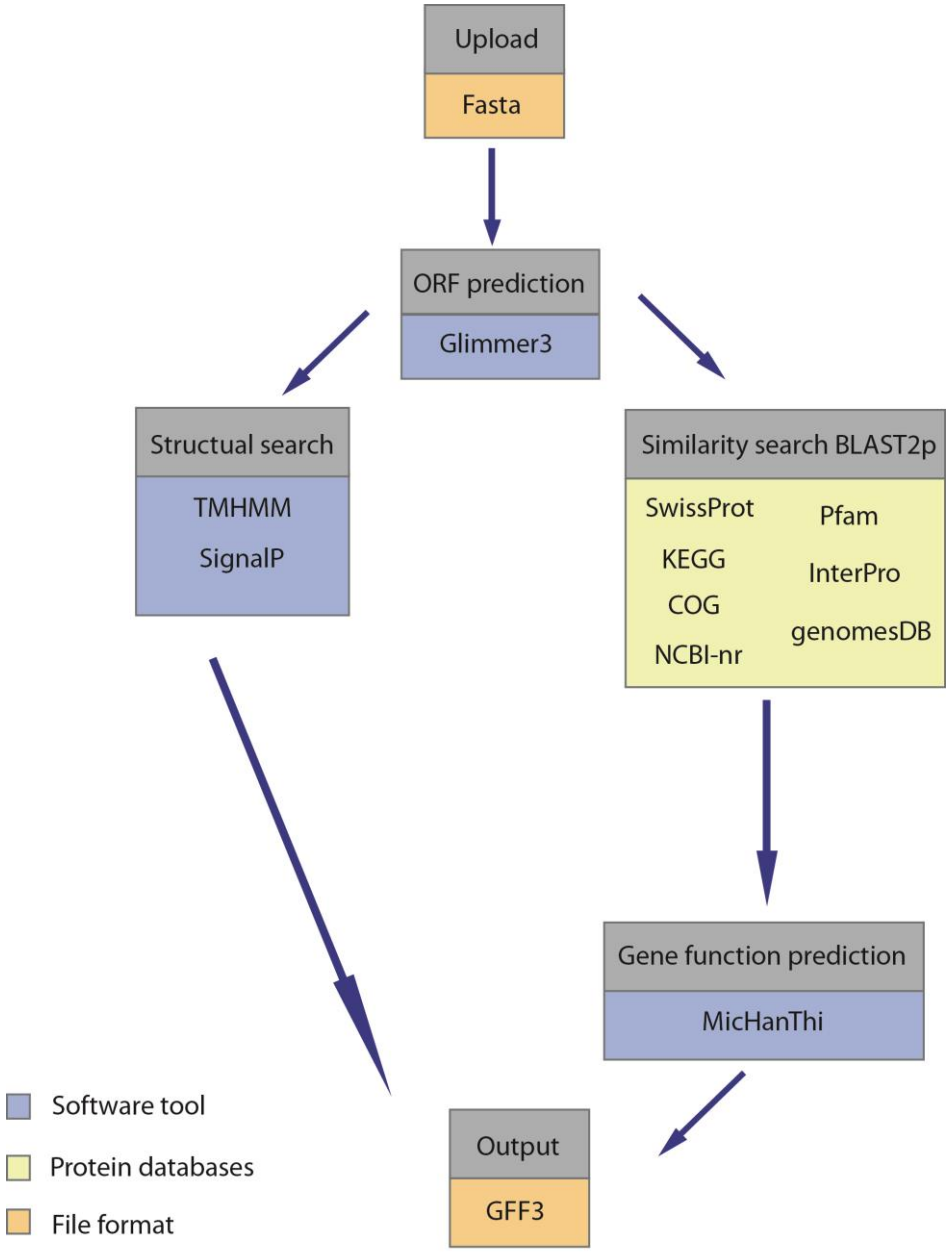

FIGURE VII FLOW CHART OF THE PROKARYOTIC GENOME ANNOTATION PIPELINE: EACH STEP (GREY) INVOLVES DIFFERENT SOFTWARE TOOLS (BLUE) AND PROTEIN DATABASES (YELLOW). FILE FORMATS ARE SHOWN IN ORANGE.

GenDB 2.2.1 is an open source genome annotation system for prokaryotic genomes relying on a relational database backend. The system provides a flexible and transparent infrastructure which allows modification to meet users requirement(Meyer, 2003).
Input: The genome sequence hast to be quality checked and assembled by the user before the MB3-Prokaryotic genome pipeline can be used. The pipeline accepts FASTA files as input format available on FTP or WEB. The URL is used to download the genome data prior to analysis.
ORF prediction by Glimmer3: Glimmer3 is a ORF prediction tool widely used in the field of prokaryotic genomics. It uses interpolated Markov models (IMMs) to identify the coding regions and distinguish them from non-coding DNA(Delcher *et al.*, 2007). The Markov model is trained by long non-overlapping ORFs of the given genome sequence itself, but only ORFs with entropy distance score less than 1.15 are considered. This is achieved by the long-orf

program integrated in Glimmer3. This newly produced Markov model is then used to predict all potential ORFs in the genome with Glimmer3.

Similarity search by BLAST2P: The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences(Altschul, 1997). Here the protein sequences of the predicted ORFs are searched against several protein databases. The settings are on default and the e-value cut-off is set to 0.01. Also, the size of the different databases is taken into account with the setting -z to make the e-values comparable.

Protein databases:

NCBI-nr database is compiled by the NCBI (National Center for Biotechnology Information, US) as a protein database for Blast searches. It contains non-identical sequences from GenBank CDS translations.

PDB, Swiss-Prot, PIR, and PRF: SWISS-PROT is a protein sequence and knowledge database that is valued for its high quality annotation, the usage of standardized nomenclature, direct links to specialized databases and minimal redundancy (Boeckmann, 2003).

COG: The database of Clusters of Orthologous Groups of proteins (COGs) is an attempt on a phylogenetic classification of the proteins encoded in 21 complete genomes of bacteria, archaea and eukaryotes. The COGs database has been designed in order to classify proteins from completely sequenced genomes on the basis of the orthology concept (Tatusov, 2000).

KEGG is a database resource for protein interaction knowledge and chemical reactions for various cellular processes (Kanehisa, 2002).

GenomesDB is a custom designed relational database, which includes a Java interface for maintenance. It is built from the proteome FASTA files obtained by the NCBI Reference Sequences database (RefSeq) for all fully sequenced bacterial and archaeal genomes. In contrast to the general-purpose database, NCBI-nr, the focus of genomesDB is to provide manually curated phylogenetic affiliations, plus as much additional contextual information as possible (Richter *et al.*, 2008).

Pfam is a database of manually currated protein families which are represented by multiple sequence alignments and hidden Markov models. Families are sets of protein regions that share a significant degree of sequence similarity, thereby suggesting homology (Finn *et al.*, 2010).

InterPro integrates together predictive models or 'signatures' representing protein domains, families and functional sites from multiple, diverse source databases: Gene3D, PANTHER, Pfam, PIRSF, PRINTS, ProDom, PROSITE, SMART, SUPERFAMILY and TIGRFAMs (Hunter *et al.*, 2009).

SignalP3.0: A signal peptide is an N-terminal signal that directs the protein across the plasma membrane. SignalP predicts these signal peptides in genomes based on neural networks which were trained on separate sets of prokaryotic and eukaryotic sequences (Bendtsen *et al.*, 2004). SignalP3.0 is run on the settings of gram negative bacteria, truncates each sequence to a maximum of 80 residues and gives a summary output.

TMHMM predicts regions in a protein with transmembrane helices, based on hidden Markov models. Proteins containing transmembrane helices usually both sides of the cell membrane. This information can help derive better understanding of the function of a protein (Krogh *et al.*, 2001). TMHMM runs on default settings.

MicHanThi is functional annotation software developed by the Microbial Genomics and Bioinformatics research group of the MPI Bremen and used in the pipeline as a tool for the prediction of gene functions. It was designed to facilitate the genome annotation process by providing a rapid, high quality prediction of gene functions. It clearly out-performed the human annotator in terms of accuracy and reproducibility of annotations during the initial

evaluation phase. Today, it is used in most genome annotation projects supervised by the Microbial Genomics Group at the institute. It is also applied by Ribocon in their industry grade annotation projects (Quast, 2007).

Output:

GFF3 is a flat tab-delimited file format. The first line of the file is a comment that identifies the file format and version. This is followed by a series of nine data lines representing sequence id, source, type, start, end, score, strand, phase and attributes. The simplicity of the file format gives the ability to modify the file in a simple text editor or with simple shell commands. http://www.sequenceontology.org/resources/gff3.html

JCoast: For manual inspection of the annotation result, we propose JCoast as an analysis and search tool. JCoast is a free open source software based on Java and runs with GenDB, developed by the Microbial Genomics and Bioinformatics research group of the MPI Bremen. It offers an easy to use genome browser which gives the ability to search easily trough all similarity search results. It also provides search tools and statistical methods especially helpful in relation to comparative genomics

Interoperability

WP5 agreed on a common output format GFF3. The strategy of all pipelines returning the same output format will make future downstream analyses easier to conduct.

Prokaryote pipeline references

Altschul, S. 1997. “Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs.” Nucleic Acids Research 25 (17) (September): 3389–3402. doi:10.1093/nar/25.17.3389.

Boeckmann, B. 2003. “The SWISS-PROT Protein Knowledgebase and Its Supplement TrEMBL in 2003.” Nucleic Acids Research 31 (1) (January): 365–370. doi:10.1093/nar/gkg095.

Delcher, A. L., K. A. Bratke, E. C. Powers, and S. L. Salzberg. 2007. “Identifying Bacterial Genes and Endosymbiont DNA with Glimmer.” Bioinformatics 23 (6) (January 19): 673–679. doi:10.1093/bioinformatics/btm009.

Bendtsen, Jannick Dyrløv, Henrik Nielsen, Gunnar von Heijne, and Søren Brunak. 2004. “Improved Prediction of Signal Peptides: SignalP 3.0.” Journal of Molecular Biology 340 (4) (July): 783–795. doi:10.1016/j.jmb.2004.05.028.

Emanuelsson, Olof, Søren Brunak, Gunnar von Heijne, and Henrik Nielsen. 2007. “Locating Proteins in the Cell Using TargetP, SignalP and Related Tools.” Nature Protocols 2 (4): 953–971. doi:10.1038/nprot.2007.131.

Finn, Robert D, Jaina Mistry, John Tate, Penny Coggill, Andreas Heger, Joanne E Pollington, O Luke Gavin, et al. 2010. “The Pfam Protein Families Database.” Nucleic Acids Research 38 (Database issue) (January): D211–222. doi:10.1093/nar/gkp985.

Hunter, S., R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, et al. 2009. “InterPro: The Integrative Protein Signature Database.” Nucleic Acids Research 37 (Database) (January): D211–D215. doi:10.1093/nar/gkn785.

Hunter, S., P. Jones, A. Mitchell, R. Apweiler, T. K. Attwood, A. Bateman, T. Bernard, et al. 2011. “InterPro in 2011: New Developments in the Family and Domain Prediction Database.” Nucleic Acids Research 40 (D1) (November 16): D306–D312. doi:10.1093/nar/gkr948.

Kanehisa, Minoru. 2002. “The KEGG Database.” In Novartis Foundation Symposia, 247:91–103. Chichester, UK: John Wiley & Sons, Ltd. http://doi.wiley.com/10.1002/0470857897.ch8.

Krogh, A, B Larsson, G von Heijne, and E L Sonnhammer. 2001. “Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete Genomes.” Journal of Molecular Biology 305 (3) (January 19): 567–580. doi:10.1006/jmbi.2000.4315.

Meyer, F. 2003. “GenDB--an Open Source Genome Annotation System for Prokaryote Genomes.” Nucleic Acids Research 31 (8) (April): 2187–2195. doi:10.1093/nar/gkg312.

Quast, Christian. 2007. “MicHanThi Design and Implementation of a System for the Prediction of Gene Functions in Genome Annotation Projects”. Dipolma thesis, University Bremen.

Richter, Michael, Thierry Lombardot, Ivaylo Kostadinov, Renzo Kottmann, Melissa Duhaime, Jörg Peplies, and Frank Glöckner. 2008. “JCoast – A Biologist-Centric Software Tool for Data Mining and Comparison of Prokaryotic (meta)genomes.” BMC Bioinformatics 9 (1): 177. doi:10.1186/1471-2105-9-177.

Tatusov, R. L. 2000. “The COG Database: A Tool for Genome-Scale Analysis of Protein Functions and Evolution.” Nucleic Acids Research 28 (1) (January): 33–36. doi:10.1093/nar/28.1.33.