Marine Microbial Biodiversity, Bioinformatics & Biotechnology

Grant agreement n°287589

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

Deliverable 6.1 Re-annotation of reference viral and giral genomes

Exploring Ecosystems Biology

Version: 2.0, 28 December 2012
Circulated to: Chris Bowler, Frank Oliver Glöckner (14 Dec. 12), Renzo Kottman, Sarah Hunter, Silvia Gonzalez-Acinas, Dolors Vaqué (21 Dec. 12).
Approved by: Pascal Hingamp, Hiroyuki Ogata (14 Dec. 12)

Expected Submission Date: 31.12.2012
Actual submission Date: 03.01.2013
Lead Party for Deliverable: Partner 6: CNRS, IGS Marseille

Mail: emilie.villar@igs.cnrs-mrs.fr **Tel.:+ 33 4 91 82 54 36**

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)	

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.

Summary

Although viruses have a major role in marine ecosystems, their large diversity makes them still poorly known. The use of metagenomics seems promising to enrich our knowledge on virus genetic diversity. Nevertheless, information about virus is dispersed in several databases and is occasionally inaccurate.
In this deliverable, we tried to extend viral sequence annotations, and to compile virus related information in one database: the Micro B3 Viral genomes DB. Starting from RefSeq Viral genomes, we searched for intergenic ORFs (Ig ORFs) which could have been missed during the original annotation. This allowed us to find 6967 IgORFs which share similarities with other previously described proteins.
In RefSeq, information about viral hosts were found to be limited. Therefore, we combined host information from several databases. In the resulting Micro B3 Viral genomes DB, host information was attributed for 96% of the recorded 4215 genomes.
This new database facilitates annotation of metagenomic datasets and biological interpretation of the results, especially for co-occurrence studies.

1. Introduction

Each day, marine viruses kill 20 to 40% of the oceanic plankton. For prokaryotes, virus dependent turnover is equivalent to the microbial mortality caused by grazing (Suttle, 2007). Thus, marine viruses play an important role in marine ecosystems, regulating population structures and thus affecting biogeochemical cycles (Fuhrman, 1999).

Marine viruses are the most abundant biological agent in the oceans, with an average abundance of $3x10^9$/L (Suttle, 2005), which represents approximately 15-fold the bacterial abundance (Suttle, 2007). Taking into account their carbon composition and their size, their biomass in the oceans is estimated to represent in the order of 200 Mt of carbon (equivalent to the carbon in 75 million blue whales).

In addition to their large population size, viruses probably have an ancient origin and are subject to high rates of mutation and recombination, leading to a high degree of diversity. This large diversity is observable at different levels: virion morphology, virus life cycle, genome molecule type and also their specific genetic content as shown by the high proportion of ORFans in viral genomes (Ogata and Claverie, 2007).

Environmental virus samples are ideal candidates for metagenomic analyses, with the small genome sizes of viruses facilitating the reconstruction of their genomes (compared to bacterial or eukaryotic genomes). Nevertheless, viral diversity is poorly represented in the existing reference databases: on average, less than 30-40% of sequences from viral community samples display significant BLAST hits within the GenBank non-redundant database (Edwards and Rohwer, 2005).

Information about viral diversity (both genotypic and phenotypic) is dispersed in different databases (mainly RefSeq, Uniprot/SwissProt, ICTV), and is occasionally inaccurate (some genomes missing annotations for essential core genes). There is therefore a critical need to consolidate this knowledge in a unique database, which will allow comprehensive annotation of the viral subset of metagenomic datasets.

2. Generating Micro B3 Viral genomes DB

2.1 Addition of new ORFs

2.1.1 Using RefSeq as a template

After comparison of available public reference databases, we confirmed that by far the most complete viral genome database was RefSeq Viral genomes (available at ftp://ftp.ncbi.nlm.nih.gov/refseq/release/viral/). In the 10/07/12 release, 4193 viral genomes were annotated and described in 4 text files:

- viral.1.protein.faa.gz: a fasta file containing coded proteins sequences.
- viral.1.genomic.fna.gz: a fasta file containing genome sequences.
- viral.1.protein.gpff.gz containing several coded proteins information.
- viral.1.genomic.gbff.gz containing several genomes information.

We added 22 already annotated genomes which were not present in the database and are important for marine ecosystem studies:

- Sputnik virophage 2 (JN603369.1)
- Ostreococcus tauri virus RT-2011 (JN225873.1)
- 7 strains of Emiliania huxleyi virus (JF974318.1, JF974310.1, JF974311.1, JF974317.1, HQ634145.1, JF974291.1, JF974290.1
- 2 strains of Micromonas pusilla virus (JF974320.1, HQ633072.1)
- 2 strains of Phaeocystis globosa virus (HQ634144.1, HQ634147.1)
- 8 strains of Organic Lake phycodnavirus (HQ704807.1, HQ704805.1, HQ704803.1, HQ704801.1, HQ704808.1, HQ704806.1, HQ704804.1, HQ704802.1)
- Paramecium bursaria chlorella virus (DQ491001.1)

This first compendium (RefSeq+22) contained 114,008 proteins and 11,0876 CDS and represented 103 virus families. Half of the viruses are DNA viruses, the others being RNA viruses.

2.1.2 Searching for new ORFs

Close inspection of the Refseq+22 compendium revealed some critical missing annotations. As an example, DNA polymerase was missing in the genome of Staphylococcus phage G1 (NC_007066, Caudovirales, Myoviridae), whilst this enzyme has been shown to be conserved in the T7-phages replication modules (Weigel and Seitz, 2006). Similarly, the capsid protein pVIII sequence is missing in the genome of the Human adenovirus D (AC_000006.1, Adenoviridae), whilst B. Harrach (in King et al., 2011) showed that this enzyme is conserved in all Adenoviridae genera. We therefore decided to search if *bona fide* protein coding genes

have been missed in intergenic sequences of RefSeq+22.

For the bioinformatics analyses, we mainly used EMBOSS applications found here: http://emboss.sourceforge.net/docs/

After extraction of intergenic sequences using maskfeat, we used getorf to search for intergenic open reading frames (IgORFs) with a specified minimum size of 90 nucleotides between two STOP codons (taking into account the genetic code specific for each virus). We obtained 188,818 putative IgORFs.

To focus on the most probable IgORFs, we searched IgORF similarities with sequences contained in the following databases (using blastp for nr, env and RefSeq Viral proteins, and using rpsblast for CDD):

nr : All non-redundant GenBank CDS translations+RefSeq Proteins+PDB+SwissProt+PIR+PRF.

(ftp://ftp.ncbi.nlm.nih.gov/blast/db/)

env: Protein sequences from large environmental sequencing projects like Sargasso Sea, Acid Mine Drainage, etc. Its entries are NOT present in nr database.

(ftp://ftp.ncbi.nlm.nih.gov/blast/db/)

RefSeq Viral proteins

(ftp://ftp.ncbi.nlm.nih.gov/refseq/release/viral/)

CDD: Conserved Domain Database, a collection of well-annotated multiple sequence alignment models for ancient domains and full-length proteins.

(ftp://ftp.ncbi.nih.gov/pub/mmdb/cdd/)

6967 IgORFs displaying a significant hit (e-value < 10^{-10}) with sequences of at least one of the above databases were selected, representing 1042 distinct genomes. These new IgORFs were added to the previously known proteins from RefSeq+22, laying the foundation of the Micro B3 virus DB.

2.2 Extending and formatting the host information

Unfortunately, in RefSeq Viral Genomes, host information is displayed as a feature encoded as free text. This leads to typing errors problems (e.g. *Abelmoschus esculentus* can be found as *Abelmoscus esculentus, Abelmoshus esculentus,* or *Albelmoschus esculentus)* and sometimes vulgar names are used (such as plantain for *Musa × paradisiaca*).

Where available, we used specific host information found in RefSeq and Uniprot databases, and in last resort general host information found in ICTV report (King et al., 2011) to provide a complete set of host annotations in the Micro B3 Viral database.

Since natural hosts are provided as a list for each virus with at least one protein sequence entry in UniProtKB/Swiss-Prot (http://www.uniprot.org/help/taxonomy), we first reduced this list to a last common ancestor, and retrieved its TaxID in NCBI (http://www.ncbi.nlm.nih.gov.gate1.inist.fr/taxonomy).

If host information was not available in UniProtKB/Swiss-Prot, we used the host information found in Refseq features, and retrieved its TaxID from the NCBI taxonomy database.

If no specific host was described for the considered virus in the two previous databases, we annotated the virus genomes with the general host contained in the ICTV report (provided in the ICTV report for each virus family).

Important note: at two stages of the host annotation process, manual curation of the data was required. Indeed, firstly in Uniprot viral vectors are sometimes included in host annotations (leading to uninformative last common ancestor of the host list, e.g. insect + plant). Secondly, typos in host names provided by RefSeq had to be corrected manually.

2.3 Summarizing information by genome

For each genome described in the MicroB3 virus database, we summarized essential viral genome characteristics:

Genomic information: Length, Type, Topology, Segments, GC content, Genetic code

Taxonomic information: Full taxonomic assignment, Family.

CDS information: Number of CDS, Mean, Max and Min length.

Intergenic ORFs (Ig ORFs): Number of IgORFs, Mean, Max and Min length.

Known virus hosts: Specific host, General host, Source.

3. Description of Micro B3 Viral genomes DB

3.1 Summary of the database:

Genome sizes (bp):

Min.	1st Quartile	Median	Mean	3rd Quartile	Max.
200	2625	5284	23490	14000	1259000

Genome types:

cRNA	DNA	ds-RNA	ms-DNA	RNA	ss-DNA	ss-RNA
72	1598	315	2	797	567	864

Genome topologies:

circular	linear
1315	2900

Genomes constituted by several segments:

multiple	single
1695	2520

GC content:

Min.	1st Quartile	Median	Mean	3rd Quartile	Max.
0.18	0.39	0.43	0.44	0.49	0.76

The 10 most represented families:

Reoviridae	479
Geminiviridae	425
Siphoviridae	298
Polydnaviridae	230
Satellites	163
Myoviridae	152
Podoviridae	112
Papillomaviridae	109
Potyviridae	103
Bunyaviridae	90

Number of CDS per genome:

Min.	1st Quartile	Median	Mean	3rd Quartile	Max.
0	1	3	26.3	8	1120

Mean length of CDS (bp):

Min.	1st Quartile	Median	Mean	3rd Quartile	Max.
0	629	969.8	1739	2078	17480

3.2 Added IgORFs

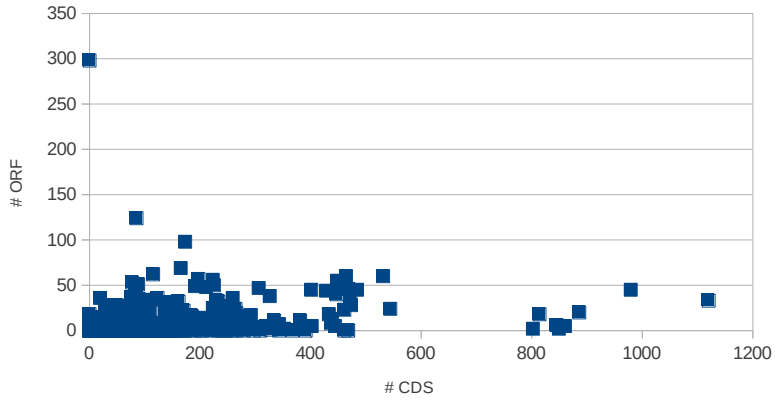
3.2.1. Distribution of IgORFs across virus genomes and virus families

6967 IgORFs displaying a significant hit (e-value < 10^{-10}) with sequences of at least one of the following database were selected, representing 1042 different genomes.

Number of IgORFs per genome:

Min.	1st Quartile	Median	Mean	3rd Quartile	Max.
0	0	1	3.5	3	298

Distribution of IgORFs number in function of the previously known CDS:



For a large part of the genomes, the proportion of missing (newly added) IgORFs was quite anecdotic (highly variable), but for 268 genomes new IgORFs represented more than 25% of the previously annotated CDSs. The genomes which presented large numbers of new IgORFs were:

NC_011335 (Diadromus pulchellus ascovirus 4a): 298

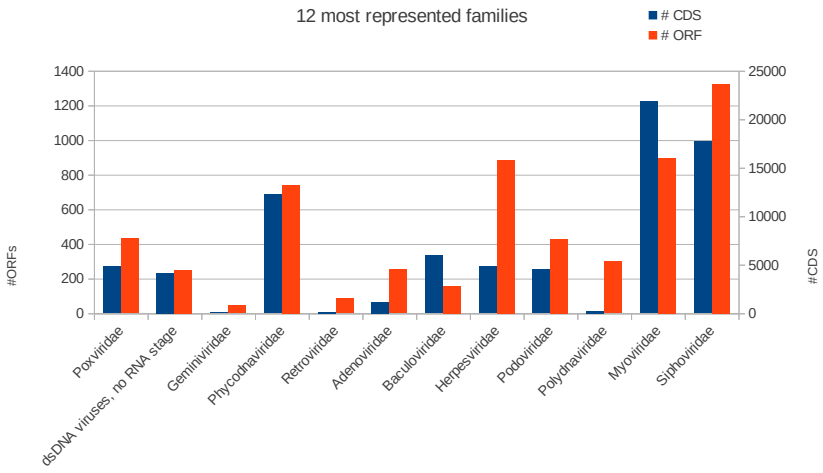
NC_002229 (Gallid herpesvirus 2): 124

NC_004105 (Ectromelia virus): 98

NC_006273 (Human herpesvirus 5): 69

NC_005880 (Staphylococcus phage K): 62

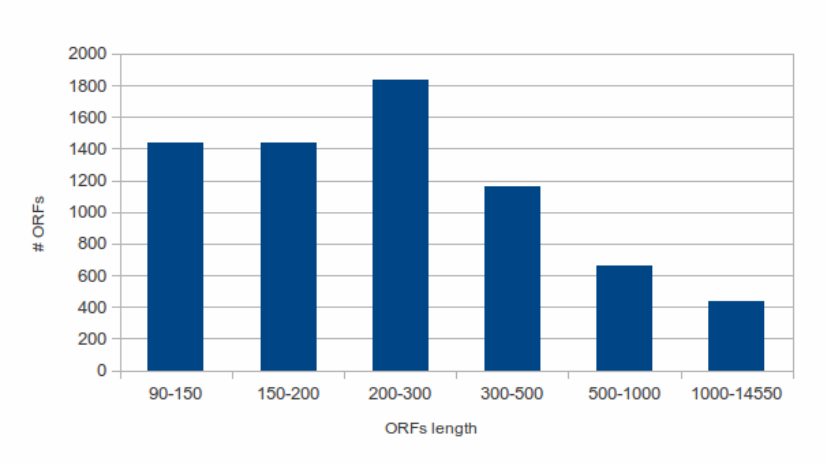
Compiled by virus families, the number of new IgORFs are:



For several families we found a large proportion of new IgORFs compared to previously known CDS, most notably Adenoviridae, Herpesviridae and Polydnaviridae.

3.2.2 Distribution of IgORFs sizes

Distribution of IgORFs length (bp):



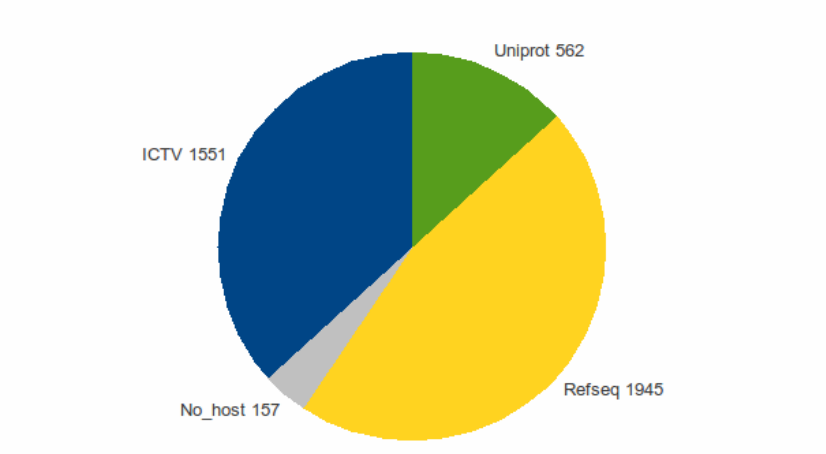
A large majority of IgORFs are small, their length being inferior to 500 nucleotides (5875 IgORFs) whilst the median length of known CDS is around 1000bp.

Mean length of IgORFs (bp):

Min.	1st Quartile	Median	Mean	3rd Quartile	Max.
90	162	225	365	360	14550

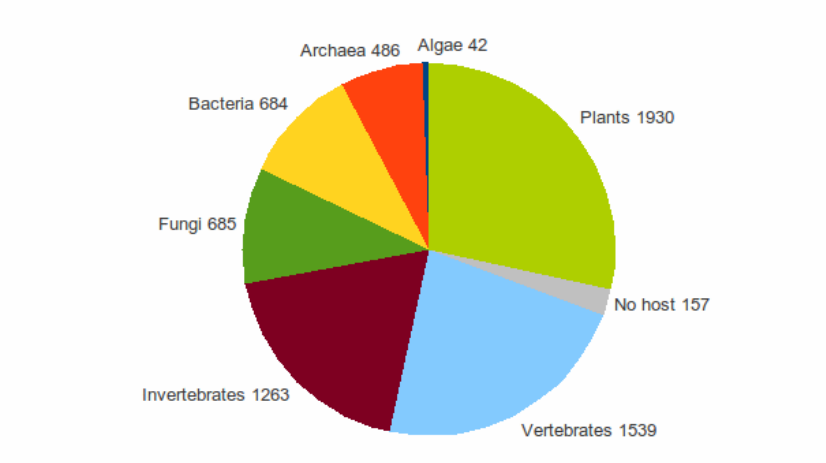
3.3 Host information

Using RefSeq, Uniprot/SwissprotKB and ICTV database, we could add host information to 96% of the viral genomes:



It must be noted that host information originating from the ICTV report (37%) are taxonomically less specific than for RefSeq and Uniprot sources (59%).

79% of the database's genomes came from eukaryotic viruses:



4. Technical considerations

Information are provided here: https://projects.mpi-bremen.de/micro-b3/trac/wiki/WorkPackages/Wp6

The MicroB3_Viral_genomes_DB contains 5 files:
MicroB3_Viral_genomes.fna
MicroB3_Viral_nucleotides.fna
MicroB3_Viral_proteins.faa
MicroB3_Viral_summary
readme

Reference list

Edwards, R.A., and Rohwer, F. (2005). Viral metagenomics. Nature Reviews Microbiology 3, 504–510.

Fuhrman, J.A. (1999). Marine viruses and their biogeochemical and ecological effects. Nature 399, 541–548.

King, A.M., Lefkowitz, E., Adams, M.J., and Carstens, E.B. (2011). Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses (Elsevier).

Ogata, H., and Claverie, J.-M. (2007). Unique genes in giant viruses: regular substitution pattern and anomalously short size. Genome Res. 17, 1353–1361.

Suttle, C.A. (2007). Marine viruses — major players in the global ecosystem. Nature Reviews Microbiology 5, 801–812.

Weigel, C., and Seitz, H. (2006). Bacteriophage replication modules. FEMS Microbiology Reviews 30, 321–381.