Identifying distant homologous viral sequences in metagenomes using protein structure information

M. Boccara M. Carpentier J. Chomilier F. Coste
C. Galiez J. Pothier A. Veluchamy

May 26, 2014

Marine viruses have a huge impact on marine life: it is estimated that they daily kill about 20% of the ocean biomass. However, marine viruses are poorly known and characterized. One lever that could help to give better knowledge of marine viruses is marine metagenomics [Sut07].
Between 2009 and 2011, the TARA Ocean [EC11] expedition sampled oceans at different depths, for different sizes of organisms. This project gave rise to a huge amount of scientific data, and specifically to metagenomic sequences.
Standard annotation tools such as MEGAN [DAJS07], PhyloPythiaS [PRH$^+$11] and Phymm [AS09], are either sequence-homology or composition based and they largely depends on the already available reference sequences. As viral sequences are highly divergent, the annotation of such data is very poor (in previous marine viral metagenomic studies, over 90% of sequences have no recognized homology [Sut07]).
It is therefore necessary to change the paradigm of current methods to improve the detection of viral sequences in metagenomes using other sources of information such as protein structure. Our goals are both to develop new methods and databases and to contribute measuring the impact of virus over marine ecosystems. We want first to adapt recently developed methods combining sequence and structure information.
A family of proteins, the capsid proteins, is present in every all viruses. Due to geometrical constrains on the viral capsid, these proteins are limited to a very small number of folds [KB11]. Moreover these folds are characteristic of the shape of the viral capsid [SC13]. Thus, identifying the fold of a protein present in a metagenomic dataset could help to identify the type of virus it is associated to.

In the PEPS VAG project, we aim at developing and adapting methods that allow designing robust sequential protein patterns taking advantage of structural information and applying them on a viral fraction of TARA Ocean metagenomic data. Aside from the novel developments, existing fold recognition methods like FROST will be adapted to our goals. FROST has been developed in the team of JF Gibrat with whom we collaborated [JRJ$^+$02]. Given a protein structure library, it assigns its most probable fold to a given sequence. We will build a library of capsid folds by comparing the protein structures of similar capsids. We will also add the information of MIRs (Most Interacting Residues) developed by one of the participating teams [PEB$^+$04] to improve its performance. MIRs

are positions dispersed along the sequence, predicted from folding simulations on a lattice. They allow to give a signature of a fold, because they statistically correspond to residues compulsory on a structural point of view.

We are also developing a novel methodology in the PEPS VAG. The idea is to extract portions of proteins – called contact motifs – capturing strong structural conservation along with relevant sequential signals. We are heading to develop short signatures at the sequential level of characteristic structures, such as capsids, that can be used to find their occurrences on sequenced data. Our method identifies characteristic structural motifs of capsid proteins of a same fold from which we deduce a characteristic signal of the underlying sequence. We present here the first experiment we made on the viral fraction of station 23 of TARA Oceans at the Deep Chlorophyll Maximum depth.

References

[AS09] Brady A and Salzberg SL. Phymm and phymmbl: metagenomic phylogenetic classification with interpolated markov models. *Nat Methods.*, 2009.

[DAJS07] Huson DH, Auch AF, Qi J, and Schuster SC. Megan analysis of metagenomic data. *Genome Res.*, 2007.

[EC11] Karsenti E and Tara Oceans Consortium. A holistic approach to marine eco-systems biology. *PLoS Biol.*, 2011.

[JRJ+02] Pley J., Andonov R., Gibrat JF., Marin A., and Poirriez V. Parallélisations d'une méthode de reconnaissance de repliements de protéines (frost). *JOBIM*, 2002.

[KB11] Mart Krupovic and Dennis H Bamford. Double-stranded dna viruses: 20 families and only five different architectural principles for virion assembly. *Current Opinion in Virology*, 2011.

[PEB+04] N. Papandreou, E. Eliopoulos, I Berezovsky, A. Lopes, E. Eliopoulos, and J. Chomilier. Universal positions in globular proteins : observation to simulation. *Eur. J. Biochem.*, 2004.

[PRH+11] Patil, Kaustubh R., Haider, Peter, Pope, Phillip B., Turnbaugh, Peter J., Morrison, Mark, Scheffer, Tobias, McHardy, and Alice C. Taxonomic metagenome sequence assignment with structured output models. *Nat Methods*, 2011.

[SC13] Cheng S and Brooks CL. 3rd viral capsid proteins are segregated in structural fold space. *PLoS Comput Biol*, 2013.

[Sut07] Curtis A. Suttle. Marine viruses — major players in the global ecosystem. *Nature Reviews Microbiology*, 2007.