Virology Journal

BioMed Central

Research

Open Access

Mimivirus relatives in the Sargasso sea

Elodie Ghedin[1,2] and Jean-Michel Claverie*[3]

Address: [1]Department of Parasite and Virus Genomics, The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA, [2] Department of Microbiology and Tropical Medicine, George Washington University, Washington DC, USA and [3]Structural and Genomics Information laboratory, CNRS-UPR2589, IBSM, 13402, Marseille, France; University of Mediterranee School of Medicine, 13385, Marseille, France

Email: Elodie Ghedin - eghedin@tigr.org; Jean-Michel Claverie* - Jean-Michel.Claverie@igs.cnrs-mrs.fr

* Corresponding author

Published: 16 August 2005

Virology Journal 2005, **2**:62 doi:10.1186/1743-422X-2-62

Received: 01 May 2005

Accepted: 16 August 2005

This article is available from: http://www.virologyj.com/content/2/1/62

Abstract

The discovery and genome analysis of *Acanthamoeba polyphaga Mimivirus*, the largest known DNA virus, challenged much of the accepted dogma regarding viruses. Its particle size (>400 nm), genome length (1.2 million bp) and huge gene repertoire (911 protein coding genes) all contribute to blur the established boundaries between viruses and the smallest parasitic cellular organisms. Phylogenetic analyses also suggested that the Mimivirus lineage could have emerged prior to the individualization of cellular organisms from the three established domains, triggering a debate that can only be resolved by generating and analyzing more data. The next step is then to seek some evidence that Mimivirus is not the only representative of its kind and determine where to look for new Mimiviridae. An exhaustive similarity search of all Mimivirus predicted proteins against all publicly available sequences identified many of their closest homologues among the Sargasso Sea environmental sequences. Subsequent phylogenetic analyses suggested that unknown large viruses evolutionarily closer to Mimivirus than to any presently characterized species exist in abundance in the Sargasso Sea. Their isolation and genome sequencing could prove invaluable in understanding the origin and diversity of large DNA viruses, and shed some light on the role they eventually played in the emergence of eukaryotes.

Introduction

The discovery and genome sequence analysis of Mimivirus [1,2], the largest of the Nucleo-cytoplasmic Large DNA Viruses (NCLDV), challenged much of the accepted dogma regarding viruses. Its particle size (>400 nm), genome length (1.2 million bp) and extensive gene repertoire (911 protein coding genes) all contribute to blur the established boundaries between viruses and the smallest parasitic cellular organisms such as Mycoplasma or Nanoarchea [2]. In the universal tree of life, the Mimivirus lineage appears to define a new branch, predating the emergence of all established eukaryotic kingdoms [2]. Although this result is compatible with various hypotheses implicating ancestral DNA viruses in the emergence of eukaryotes [3-5], it requires confirmation from additional data. An urgent task is thus to convince ourselves that Mimivirus is not the sole representative of its kind (i.e. a viral counterpart to the platypus) and to provide some rational guidance as to where to begin the search for eventual new *Mimiviridae*.

Mimivirus was serendipitously discovered within *Acanthamoeba polyphaga*, a free-living ubiquitous amoeba, prevalent in aquatic environments. Phylogenetic analysis of the most conserved genes common to all nucleo-cytoplasmic large double-stranded DNA viruses (NCLDV) [6]

Table 1: Matching Status of Mimivirus core genes (type 1 to 4).

ORF#	Definition	Best score in nr	Best score in DNA viruses	Best score in Sargasso Sea	Status	Reciprocal Best match
L206	Helicase III / VV D5	167-virus	167	214	Best ENV	YES
R322	DNA pol (B family) extein	207	167	238	Best ENV	YES
L437	**A32 virion packaging ATPase**	**169-virus**	**169**	**191**	**Best ENV**	**YES**
L396	VV A18 helicase	200-virus	200	187	-	
L425	Capsid protein	119-virus	117	142	Best ENV	complex
R439	Capsid protein	164-virus	159	173	Best ENV	complex
R441	Capsid protein	137-virus	147	209	Best ENV	complex
R596	E10R-Thiol oxidoreductase	104-virus	105	119	Best ENV	YES
R350	VV D6R – helicase	170-virus	170	102	-	
R400	F10L – prot. Kinase	86-virus	86	58	-	
R450	A1L-transcr factor	52-virus	47	65	Best ENV	
R339	TFII-transcr. factor	62	42	66	Best ENV	
L524	MuT-like NTP PP-hydrolase	40	38	39	-	
L323	Myristoylated virion prot. A	43	42	40	-	
R493	PCNA	92	87	154	Best ENV	YES
L312	Small Ribonucl. reduct	341	338	310	-	
R313	Large Ribonucl. reduct	766	741	740	-	
R429	**PBCV1-A494R-like**	**152-virus**	**152**	**216**	**Best ENV**	**YES**
L37	BroA, KilA-N	123-virus	124	65	-	
R382	mRNA-capping enz.	86	78	166	Best ENV	YES
L244	RNA pol. sub 2 (Rbp2)	727	416	508	-	
R501	RNA pol. sub.1 (Rpb1)	805	415	520	-	
R195	ESV128-Glutaredoxin	50	39	49	-	
R622	S/Y phosphatase	75	73	65	-	
R311	CIV193R BIR domain	68	44	51	-	
L65	Virion memb. prot	44	44	-	-	
R480	Topoisomerase II	902	717	367	-	
L221	Topoisomerase I bacterial	528	35	516	-	
R194	Topoisomerase I pox-like	188	100	145	-	
L364	SWI/SNF2 helicase	70-virus	70	72	Best ENV	YES
L4	N1R/P28 DNA binding prot	123-virus	124	72	-	
L540	Pre-mRNA helicase – splicing	256	136	214	-	
L235	RNA pol subunit5	69	38	50	-	
R354	Lambda-type exonuclease	69-virus	69	154	Best ENV	YES
R343	RNAse III	129	112	131	Best ENV	YES
R141	GDP mannose 4,6-dehydratase	294	68	252	-	
L258	Thymidine kinase	151	140	124	-	
L271	Ankyrin repeats (66 paralogs)	179	152	192	Best ENV	complex
R325	Metal-dependent hydrolase	69-virus	69	105	Best ENV	YES
L477	Cathepsin B	226	43	47	-	
R497	Thymidylate synthase	278	242	217	-	
R449	**Uncharacterized prot.**	**69-virus**	**69**	**129**	**Best ENV**	**YES**
R303	NAD-dependent DNA ligase	270-virus	270	228	-	
L805	MACRO domain	36	33	-	-	
R571	Patatin-like phospholipase	105	80	122	Best ENV	YES
R301	Uncharacterized prot.	48-virus	48	65	Best ENV	YES

positions Mimivirus as an independent lineage, roughly equidistant from the Phycodnaviridae (algal viruses) and Iridoviridae (predominantly fish viruses). Given the ecological affinity of these virus families for the marine environment, we have examined the sequence data set gathered through environmental microbial DNA sampling in the Sargasso Sea [7] to look for possible Mimivirus relatives.

Results

By comparing Mimivirus ORFs to the Sargasso Sea sequence data set and to all other publicly available

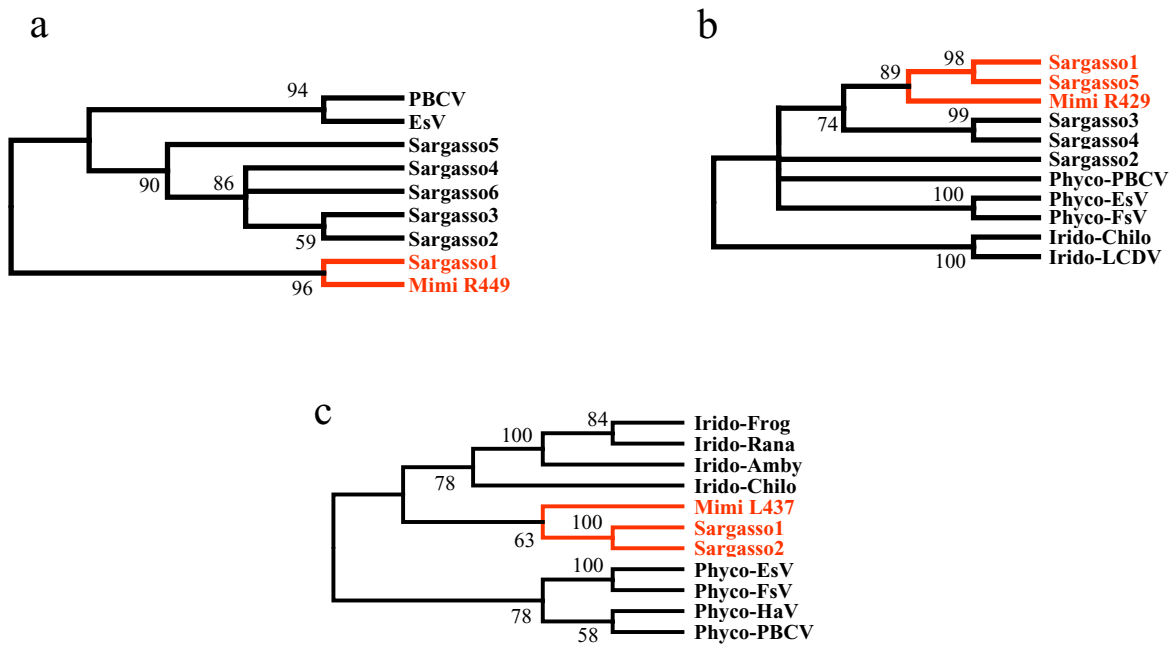

Figure 1

Phylogenetic evidence of uncharacterized Mimivirus relatives. **(a)** Neighbor-joining (NJ) clustering (see Materials and Methods) of Mimivirus R449 ORF with its best matching (≈35% identical residues) environmental homologues (noted Sargasso1 to Sargasso6 according to their decreasing similarity) and closest viral orthologues (28% identical). **(b)** NJ clustering of Mimivirus R429 ORF with its best matching (≈50% identical) environmental homologues (noted Sargasso1 to Sargasso5) and closest viral orthologues (35% identical). **(c)** NJ clustering of Mimivirus putative virion packaging ATPase L437 with its best matching (≈45% identity) environmental homologues (Sargasso1 and Sargasso2) and closest viral orthologues (34% identical). Abbreviations: Phyco: Phycodnavirus; PBCV: *Paramecium bursaria* chlorella virus 1; EsV: *Ectocarpus siliculosus* virus; FsV: Feldmannia sp. virus; HaV: Heterosigma akashiwo virus; Irido: Iridovirus; LCDV: Lymphocystis disease virus 1; Frog: Frog virus 3; Amby: *Ambystoma tigrinum stebbensi* virus; Rana: *Rana tigrina ranavirus*; Chilo: Chilo iridescent virus. Bootstrap values larger than 50% are shown. Branches with lower values were condensed.

sequences, 138 (15%) of the 911 Mimivirus ORFs were found to exhibit their closest match (Blastp E-values ranging from 10^{-74} to 10^{-4} [8]) to environmental sequences (see Additional file 1). Even before the discovery of Mimivirus, increasingly complex large double-stranded DNA viruses have been isolated, in particular from unicellular algae. The genome analysis of these Phycodnaviruses revealed a variety of genes encoding enzymes from totally unexpected metabolic pathways [9]. Mimivirus added more unexpected genes (such as translation system components [2]) to this list. As the gene repertoire of these large viruses and the gene content of cellular organisms become increasingly comparable, we have to be cautious in the interpretation of environmental/metagenomics sequence data. To focus our study on environmental organisms most likely to be viruses, we limited further analyses to Mimivirus homologues member of the NCLDV core gene sets [2,6]. These core genes are subdivided into four classes from the most (class I) to least (class IV) evolutionarily conserved [6]. Seven of 10 Mimivirus Class I core genes (L206 to R400) have their closest homologues in the Sargasso Sea data. This is also the case for 3 of 7 class II (R450-R313)core genes, 3 of the 13 class III core genes (R429-L364) and 7 of the 16 Class IV core genes (L4-R301) (Table 1). Overall, 43% of Mimivirus core genes have their closest homologues in the Sargasso Sea data set. To further assess the viral nature of these unknown microbes, we studied the phylogenetic relationships between the corresponding Mimivirus proteins, their Sargasso Sea homologues, and the closest homo-

Figure 2
Organization of four Mimivirus ORF best matching homologues in a 4.5 kb environmental sequence scaffold (approximately to scale). The three colinear Mimivirus homologues are in green. Unmatched ORF extremities are indicated by dots. The two diagonal lines indicate where the two contigs are joined on the scaffold.

logues in other NCLDVs (see Materials and Methods). Figure 1a–c exhibits three independent phylogenic trees computed using the MEGA3 software [10] for Mimivirus ORFs R449 (unknown function), R429 (unknown function) and L437 (putative virion packaging ATPase). Figure 1a shows that the closest environmental R449 homologues cluster with Mimivirus separately from the known phycodnaviruses, while other Sargasso Sea homologues cluster in a way suggesting the presence of a new clade distinct from Phycodnaviridae. The tree based on R429 and L437 (Fig. 1b,c) similarly suggests the presence of close Mimivirus relatives not belonging to the Phycodnaviridae or Iridoviridae clades.

Another piece of evidence substantiating the existence of an unknown Mimivirus relative in the Sargasso Sea is the discovery of contigs built from the data that contain multiple genes with a high degree of similarity to Mimivirus genes. A spectacular case is illustrated in Figure 2. Here, a 4.5 kb scaffold (See Materials and Method) exhibits 4 putative ORFs. When compared to the whole nr database, each of them has as a best match 4 distinct Mimivirus ORFs: thiol oxidoreductase R368 (29% identical, E-value < 10^{-9}), NTPase-like L377 (25% identical, E-value < 10^{-20}), unknown function L375 (34% identical, E-value < 10^{-30}), and DNA repair enzyme L687 (40% identical, E-value < 10^{-62}). Moreover, the gene order is conserved for three of them (R368, L375, L377). Such colinearity is rarely observed between viral genomes except for members of the same family. Unfortunately, the sequences of these genes are not conserved enough to allow the construction of informative phylogenic trees that would include other NCLDV orthologues.

As of today, genes encoding capsid proteins are among the most unequivocal genes of viral origin. Except for cases of integrated proviral genomes, no cellular homologues of viral capsid proteins have ever been found. During our study, the closest homologues of Mimivirus capsid proteins were found to be capsid protein genes of environmental origin. For example, Mimivirus capsid protein (R441) was found to be 48.5% identical to an unknown environmental sequence, when it is only 36.2% identical to the major capsid protein Vp49 of Chlorella virus CVG-1, its best match among known viruses (Figure 3). As the environmental capsid protein sequence also shares 44.5% identical residues with the CVG-1 Vp49, the corresponding uncharacterized virus appears to lie at an equal evolutionary distance from the *Mimiviridae* and the *Phycodnaviridae*.

Discussion

Our results predict that DNA viruses of 0.1 to 0.8 microns in size exist in the Sargasso Sea that are evolutionarily closer to Mimivirus than to any presently characterized species. These viruses are abundant enough to have been collected by environmental sampling. It must be noticed that a similar approach attempting to find relatives to two other unique NCLDVs, the African swine fever virus (the unique member of Asfarviridae) and the White spot syndrome virus, a major shrimp pathogen (the sole Nimaviridae), failed to provide convincing results (Claverie, data not shown). The identification of numerous Mimivirus-like sequences in the Sargasso Sea data is thus not simply the result of a large number of sequences been compared, but truly suggests that viruses from this clade are specifically abundant in the sampled marine environment. It is actually expected that many novel viruses will be encountered in natural waters in which they constitute the most abundant microrganisms [11,12]. There might be as many as 10 billion virus particles per litre of ocean surface waters [13]. Interestingly, the specialized literature abounds of descriptions of large virus-like particle associated with algae [e.g. [14-16]], or various marine protists [17,18]. With the exception of Phycodnaviruses [19-21], the genomic characterization of these viruses has not been attempted. Guided by the results presented here, their isolation and genome sequencing could prove invaluable in understanding the diversity of DNA viruses and the role they eventually played in the evolution of eukaryotes.

Materials and methods

The protocols used to collect Sargasso Sea environmental micro-organisms and generate DNA sequences from these samples has been described elsewhere [7]). The data analyzed here correspond to "bacteria-sized" organisms that have passed through 3 μm filters and been retained by 0.8 μm to 0.1 μm filters. Mimivirus-like particles (0.8–0.4 μm) belong in this range.

```
CVG1-vp49       MAGGLSQLVAYGAQDVYLTGNPQITFFKTVYRRYTNFAVESIQQTINGSV
MIMI-R441       MAGGIIQLVAYGIQDLYLTGDPQITFFKVVYRRHTNFSVESIIQNFTSVP
Sargasso1       MGGGLMQLVAYGAQDIYLTGNPQITFFKVVYRRHTNFSVESIKQTFNGTA

CVG1-vp49       GFGNKVSTQISRNGDLITDIVVEFVLTKQGPTFY---------------
MIMI-R441       DFGSTVSCTLSKSGDMINKIYVYIELPSVNVFYDESG------NLDKFKK
Sargasso1       DFGKKVSCTISRNGDLVHRIFLQTTLPAQKYDYASAGGGTVTYNSNSNMK

CVG1-vp49       ---------CAEQLLQDVELEIGGQRIDKHYADWFRMYDSLFRMD-----
MIMI-R441       ---FAWVRNIGYALIKDVSIEIGGKLIDKQYGEWMYIWSQVTNKS--DEG
Sargasso1       DGILRWINWVGEKLINYAEIEIGGQRIDKHYGEWLHIWGQLTNTASHDEG

CVG1-vp49       --NDRQNYRRM-----TDFVNDEPATAVKRFYVPLIFFFNQTPGLALPLI
MIMI-R441       LDKMIGNIPLL-----YDFSNGKP---KYSLYVPLEFWFCRNSGLSLPLV
Sargasso1       YQRMVGNIPALTTNVSTNTVAGAAEIKAQDLYVPLQFWFCRNPGLALPLI

CVG1-vp49       ALQYHEVKLYFTLAST-------------VNGITAVEGGAAVTAVAP
MIMI-R441       ALSSSEVKITISFRSAEECYRIGPTHSIEIMEDIVPFEFGDYIEQKIG
Sargasso1       ALQYHEVKINIEFEEL-------------KNLFIAQEKTTAATAVTN
```

Figure 3
Partial 3-way alignment (N-terminus region) of Mimivirus capsid protein (R441) with it best matching homologues in the NR and Environmental sequence databases. The Mimivirus R441 protein shares 83/229 (36.2%) identical residues (colored in red or blue) with the major capsid protein Vp49 of Chlorella virus CVG-1 and 111/229 (48.5%) identical residues (indicated in red or green) with the N-terminus of a capsid protein from an unknown large virus sampled from the Sargasso Sea (Accession: EAD00518). On the other hand, the CVG-1 Vp49 and the Sargasso Sea sequence share 44.5% identical residues. By comparison, the CVG-1 Vp49 protein share 72% of identical residue with PBCV-1 Vp54, its best matching homologue among known phycodnaviruses.

Database similarity searches were performed using the Blast suite of programs [8] (default options) as implemented on the http://www.giantvirus.org web server and as implemented at The Institute for Genomic Research. Final similarity searches were performed on the non-redundant peptide sequence databases (nr) and environmental data (env-nr) downloaded from the National Institute for Biotechnology Information ftp server ftp://ftp.ncbi.nlm.nih.gov/blast/db/ on March 14, 2005. To avoid missing potential better matches with annotated virus ORFs, all Mimivirus ORFs exhibiting a best match (blosum62 scoring scheme) in env-nr were also searched against all DNA virus genomes using TblastN (peptide query against translated nucleotide sequence). The comprehensive list of Mimivirus ORFs exhibiting a best match in the env-nr database is given in Additional file: 1. Phylogenetic analyses were conducted using MEGA version 3.0 [10] (option: Neighbor joining, 250 pseudo-replicates, and gaps handled by pairwise deletion). Tree branches were condensed for bootstrap values <50%.

Only Mimivirus ORFs with best matching homologues in DNA viruses and belonging to the nucleo-cytoplasmic large DNA virus core gene set (2, 6) were analyzed in detail. These ORFs (and matching status) are listed in Table 1. Phylogenetic analyses were limited to viral homologues and environmental sequences exhibiting a reciprocal best match relationship with the corresponding Mimivirus ORF (putative orthologues) (YES in the rightmost column). The three cases (red lines in Table 1) exhibiting the best bootstrap values are shown in Figure 1. Cases of complex relationships, for instance due to the presence of many paralogues (e.g. capsid proteins), are also indicated. These cases of non-reciprocal best matches are frequent (i.e. the closest homologue of a Mimivirus ORFs being an environmental sequence, but the latter

sequence exhibiting a better match with a different ORF in the nr database).

Two environmental sampling contigs – contig IBEA_CTG_1979672 (AACY01022731, GI:44566181) and contig IBEA_CTG_1979673 (AACY01022732, GI:44566179) – are linked in a 4,465 bp scaffold (scaffold IBEA_SCF = 2208413) found to contain four ORFs with strong matches to Mimivirus peptides (R368, L377, L375, and L687). The three colinear ORFs (R368, L377, L375) are found on one contig while the orthologue to Mimivirus ORF L687 is found in the second contig. It is conceivable that the lack of colinearity for this fourth ORF is due to an assembly error.

Additional material

Additional file 1

List of Mimivirus ORFs exhibiting a best match in the env-nr database
Click here for file
[http://www.biomedcentral.com/content/supplementary/1743-422X-2-62-S1.pdf]

Acknowledgements

We are indebted to James van Etten for pointing out some ancient observations of very large virus-like particles in algae and marine protists. We thank Stéphane Audic for his help with the http://www.giantvirus.org server and Hiroyuki Ogata and Vish Nene for reading the manuscript. This work was supported by internal funding from TIGR, CNRS, and the French National Genopole Network.

References

1. La Scola B, Audic S, Robert C, Jungang L, de Lamballerie X, *et al.*: **A giant virus in amoebae.** *Science* 2003, **299:**2033.
2. Raoult D, Audic S, Robert C, Abergel C, Renesto P, *et al.*: **The 1.2-megabase genome sequence of Mimivirus.** *Science* 2004, **306:**1344-1350.
3. Villarreal LP, DeFilippis VR: **A hypothesis for DNA viruses as the origin of eukaryotic replication proteins.** *J Virol* 2000, **74:**7079-7084.
4. Bell PJ: **Viral eukaryogenesis: was the ancestor of the nucleus a complex DNA virus?** *J Mol Evol* 2001, **53:**251-256.
5. Takemura M: **Poxviruses and the origin of the eukaryotic nucleus.** *J Mol Evol* 2001, **52:**419-425.
6. Iyer LM, Aravind L, Koonin EV: **Common origin of four diverse families of large eukaryotic DNA viruses.** *J Virol* 2001, **75:**11720-11734.
7. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, *et al.*: **Environmental genome shotgun sequencing of the Sargasso Sea.** *Science* 2004, **304:**66-74.
8. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, *et al.*: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
9. Van Etten JL: **Unusual life style of giant chlorella viruses.** *Annu Rev Genet* 2003, **37:**153-195.
10. Kumar S, Tamura K, Nei M: **MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment.** *Brief Bioinform* 2004, **5:**150-163.
11. Wommack KE, Colwell RR: **Virioplankton: viruses in aquatic ecosytems.** *Microbiol Mol Biol Rev* 2000, **64:**69-114.
12. Ghedin E, Fraser C: **A virus with big ambitions.** *Trends Microbiol* 2005, **13:**56-57.
13. Fuhrman JA: **Marine viruses and their biogeochemical and ecological effects.** *Nature* 1999, **399:**541-548.
14. Mattox KR, Stewart KD: **Probable virus infections in four genera of green algae.** *Can J Microbiol* 1972, **18:**1620-1621.
15. Dodds JA, Cole A: **Microscopy and biology of Uronema gigas, a filamentous eukaryotic green alga, and its associated tailed virus-like particle.** *Virology* 1980, **100:**156-165.
16. Van Etten JL, Lane LC, Meints RH: **Viruses and virus-like particles of eukaryotic algae.** *Microbiol Rev* 1991, **55:**586-620.
17. Shin W, Boo SM: **Virus-like particles in both nucleus and cytoplasm of Euglena viridis.** *Algological Studies* 1999, **95:**125-131.
18. Growing MM: **Large VLPs from vacuoles of phaeodarian radiolarians and from other marine samples.** *Marine Ecology progress Series* 1993, **101:**33-43.
19. Van Etten JL, Graves MV, Muller DG, Boland W, Delaroque N: **Phycodnaviridae – large DNA algal viruses.** *Arch Virol* 2002, **147:**1479-516.
20. Claverie JM: **Giant viruses in the oceans: the 4th algal virus workshop.** *Virol J* 2005, **2:**52.
21. Wilson WH, Schroeder DC, Allen MJ, Holden MTG, Parkhill J, *et al.*: **Complete Genome Sequence and Lytic Phase Transcription Profile of a coccolithovirus.** *Science* 2005, **309:**1090-1092.