Supervised Classification of Metatranscriptomic Reads Reveals the Existence of Light-dark Oscillations During Infection of Phytoplankton by Viruses

Enzo Acerbi¹*, Caroline Chenard²*, Stephan C. Schuster¹ and Federico M. Lauro¹,²

¹Singapore Centre on Environmental Life Sciences Engineering (SCELSE), Nanyang Technological University, 60 Nanyang Dr, 637551, Singapore
²Asian School of the Environment, Nanyang Technological University, 50 Nanyang Avenue, 639798, Singapore

* These authors contributed equally.

Keywords: Support Vector Machines, Empirical Mode Decomposition, Marine Microbial Ecology.

Abstract: In the era of next generation sequencing technologies microbial species identification is typically performed using sequence similarity and sequence phylogeny based approaches. Particularly challenging is the discrimination of closely related sequences such as auxiliary metabolic genes (AMGs) in cyanobacteria and their viruses (cyanophages). Here we developed a method which combines Support Vector Machine based classification of AMGs short fragments and Empirical Mode Decomposition of periodic features in time-series. We applied this method to investigate the transcriptional dynamics of viral infection in the ocean, using data extracted from a previously published metatranscriptome profile of a naturally occurring oceanic bacterial assemblage sampled Lagrangially over 3 days. We discovered the existence of light-dark oscillations in the expression patterns of AMGs in cyanophages which follow the harmonic diel transcription of both oxygenic phototrophic and heterotrophic members of the community. These findings suggest that viral infection might provide the link between light-dark oscillations of microbial populations in the North Pacific Subtropical Gyre.

1 INTRODUCTION

Like most of life sciences, marine microbial ecology has also been revolutionized by the advent of next generation sequencing (NGS) technologies. Taxonomical identification of DNA/RNA sequences (also referred to as fragments) is typically performed by checking for the existence of a similarity by alignment to the genomes of all the known microbial species in a brute force kind of fashion. If matches are found, phylogenetic analysis may subsequently be conducted to unravel the evolutionary relationships among the species. This similarity based identification approach has two main limitations: Firstly, the extensive number of comparison to be performed makes the task computationally expensive, which translate in high costs and long processing time. Secondly, NGS technologies can generate sequences of limited length (typically 250 bp). While this may not represent an issue with distantly related sequences, it can make the discrimination of closely related sequences (high similarity) difficult. This is the case for auxiliary metabolic genes (AMGs) in cyanobacteria and their associated viruses (also referred to as cyanophages). AMGs are encoding key metabolic functions such as photosynthesis, carbon metabolism, etc. Upon infection, viruses take over the control of the bacterial cell and keep it alive expressing the AMGs that they are carrying, for which an high similarity with the bacterial AMGs is required for a successful infection.

Being able to determine whether an AMG short fragment (fragments that does not cover the entire length of the AMG) belongs to a bacterial host or its associated viruses is an extremely valuable resource in microbial ecology. In the past, alternative methods exploiting specific genetic content have been used to characterize the origin of a gene (Sandberg et al., 2001). Recently, Tzahor et al. (Tzahor et al., 2009) used a multi-class Support Vector Machine (SVM) to rapidly classify core-photosystem-II gene and transcripts fragments coming from marine samples based on their oligonucleotide frequencies. Here, we applied a SVM-based approach to classify short fragments of two different AMGs, na-
mely psbA and phoH, retrieved from a previously published metatranscriptome profiling of multiple naturally occurring oceanic bacterial populations sampled in situ over 3 days (Ottesen et al., 2014). The short psbA and phoH fragments (with length ranging between 150 and 300bp) were classified based on their origin (viral vs host) using GC content, mono-, di-, tri-, tetra-nucleotide frequencies as discriminating features.

The outputs were further processed using Empirical Mode Decomposition (EMD) to identify the underlying frequency dynamics of transcription during viral infection in environmental cyanobacterial populations. EMD can deconvolve natural time-series data into composing frequencies also known as intrinsic mode functions (IMFs). EMD has been used successfully to analyse time series data in other scientific domains including the detection of periodically expressed genes in microarray data (Chen et al., 2014), analysis of seismic data (Han and van der Baan, 2013), electrocardiograms (Chang, 2010) and anomaly in sea-surface height (Li et al., 2012). Our data demonstrated an existence of light-dark oscillation in the expression pattern of AMGs in cyanophages suggesting that viral infection patterns might provide the dynamic coupling between light-dark oscillations of autotrophic and heterotrophic microbial populations in the North Pacific Subtropical Gyre.

2 BIOLOGICAL BACKGROUND

Marine picocyanobacteria from the genera Synechococcus and Prochlorococcus are major primary producers in the ocean at low- and mid-latitude and contribute significantly to the global carbon cycle (Hess, 2004; Partensky et al., 1999). Given that light represents the main source of energy for cyanobacteria, it also determines the tempo of carbon fixation, metabolic and physiological activity such as the timing of cell division, amino acid uptake, nitrogen fixation, photosynthesis and respiration (Ni and Zeng, 2016; Golden et al., 1997). For example, nearly half of all Prochlorococcus population in the North Pacific Subtropical Gyre demonstrated a transcriptional diel cycle (Ottesen et al., 2014). Viral infection in the oceans might also be synchronized with the light cycle. Some studies suggest a diel cycle in the number of infective viruses which could mainly be attributed to UV damage during the peak of sunlight (Wilhelm et al., 1998; Suttle and Chen, 1992). However, the timing of viral replication can also be influenced by the presence and absence of light. For example, a temporal study at a station in the Indian Ocean revealed a strong increase in viral abundance in the middle of the night (Clokie et al., 2006a). Logically, the synchronicity between light and replication is especially important for viruses infecting cyanobacteria, the cyanophages. Indeed, it was shown that light influences viral fitness as light is required for viral infection as it might influence the degradation of host genomic DNA, viral transcription or production of new progeny (Thompson et al., 2011).

In addition, cyanophages harbour AMGs which are usually homologs of host genes involved in photosynthesis and carbon metabolism pathways. These AMGs likely play a role in viral infection (Breitbart et al., 2007). For example, the psbA gene which encodes D1 protein involved in the photosystem II (PSII) reaction centre is prevalent in cyanophage genomes. PSII is particularly sensitive to photodamage causing a high turnover rate for the D1 protein. It was shown that viral psbA can highly be transcribed during infection which suggests that viral-psbA expression might maintain the photosynthetic activity of the infected host cells and therefore provide energy for cyanophage replication (Clokie et al., 2006b; Lindell et al., 2007). This is supported by the notion that during the lytic cycle of infection of Prochlorococcus MED4 by cyanophage P-SSP7 most of the psbA transcript in infected cells were from viral origin after 6hrs (Lindell et al., 2007). Similarly, the phoH gene which is involved in phosphate metabolism was previously found in both heterotroph and autotroph phages (Goldsmith et al., 2011).

Based on sequence phylogeny, viral psbA can generally be distinguished from Synechococcus and Prochlorococcus (Sullivan et al., 2006; Chenard and Suttle, 2008), while phoH genes cluster together based on their origin (i.e. autotroph bacteria, heterotroph bacteria, cyanophage, heterotroph phage and eukaryotic viruses (Goldsmith et al., 2011; Goldsmith et al., 2015)).

3 METHODS

3.1 Support Vector Machines

Support Vector Machines (SVMs) are a supervised learning algorithm in machine learning, first introduced by Vapnik (Stitson et al., 1996) and based on the principle of structural risk minimization. Given labeled data (in our case, DNA fragment whose origin is known), a SVM can be trained to individuate the optimal hyperplane separating (classifying) new examples.
3.2 Training Set Generation

Labelled DNA sequences for psbA and phoH genes are scarce in the literature. In order to generate the psbA training set for the SVM-based multiclassifier, we collected 203 psbA DNA sequences from NCBI (including those used by Tzahor S. et al. (Tzahor et al., 2009)). Sequences in the dataset were labelled as belonging to one of the following five categories: *Synechococcus* bacteria (77), *Synechococcus* virus (42), *Prochlorococcus* bacteria high-light (38), *Prochlorococcus* bacteria low-light (26) and *Prochlorococcus* virus (20). Sequences were long, having different lengths ranging from 720bp to 1083bp, with a median of 1080bp. Analogously, phoH training set was created by collecting 84 DNA sequences from NCBI. Sequences in the dataset were labelled as belonging to one of the following five categories: autotrophic host (i.e. cyanobacteria) (17), cyanophage (29), heterotrophic host (13), heterotrophic phage (20).
PhoH sequences were also long, having different lengths ranging from 603bp to 1770bp, with a median of 766.5 bp. In their work, Tzahor et al. trained the SVM classifier using full length psbA sequences and subsequently used it for classification of other full length psbA sequences (Tzahor et al. also tested their classifier on short psbA fragments, but only limited to binary classification). If the aim is to use the classifier on shorter psbA fragments, the approach of Tzahor et al. may not always give the best results. In fact, classification accuracy may be low if the training is performed on clean samples (psbA full genes, rather than fragments) and the real-world classification is performed on noisy samples (psbA fragments, i.e. sequences containing a part of the psbA gene and other base pairs not related to the psbA gene, as in metagenomic/metatranscriptomic data). For this reason, psbA and phoH training sets were generated by randomly extracting sequences of length 300bp from the original sets. In addition, to deal with the original dataset being slightly unbalanced (not equal number of instances for each class) a combination of undersampling and oversampling was applied to both psbA and phoH sequences in order to obtain balanced training sets: 100 sequences for each of the five classes (total of 500 sequences) for psbA training set and 50 sequences for each of the five classes (total of 250 sequences) for phoH training set.

3.3 Feature Generation

GC, mono-, di-, tri-, tetra-nucleotide frequencies were calculated for each sequence of the training sets (for a total of 341 features per sequence) and used as input feature vector for the SVM classifiers. When including penta- and/or hexa-nucleotide frequencies, no significant improvements in prediction accuracy were observed (data not shown).

3.4 The SVM-based Model

Although through the document it will be referred to as SVM-based classifier, the model is composed by 5 different SVM one-against-all classifiers with linear kernel, each of them trained to separate one of the five taxonomical categories (positive class) from all the remaining ones (collapsed as negative class). Associated with the prediction on whether a given sequence belongs or not to the positive class for which it was trained, each classifier returns a numeric value representing the probability estimate of the prediction. For each new sequence, the classifier returning the prediction associated with the highest probability estimate is the one determining the category of the sequence.

3.5 Model Training, Parameter Optimization and Validation

The LIBSVM toolbox (v3.22) (Chang and Lin, 2011) for MATLAB was used for the experiments. The SVM-based classifier was trained using the 500 and 250 sequences of the psbA and phoH training sets respectively. In order to assess the ability of the model to correctly assign each sequence to the respective taxonomical category, a cross validation of type leave-one-out was performed. This validating procedure iterates over all of the N sequences of a training set, each time using N-1 sequences to train the classifier and 1 sequence as a test. Performances were assessed in terms of precision, recall and f-measure, which for binary classification are defined as follows:

\[ \text{precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}} \]  
\[ \text{recall} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}} \]  
\[ F_1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \]

In statistics the recall is referred to as sensitivity and the precision as positive predicted value. For psbA, cross validation mean precision resulted to be equal to 0.95, mean recall 0.92 and mean \( F_1 \) 0.93. In this phase, the optimal value for the parameter Cost (cost of misclassification) was empirically established as being equal to 8, while the optimal probability estimate cut-off resulted to be equal to 0.3 (resulting in 3.8% of the sequences being marked as unassigned). A direct comparison of the SVM-based classifier with the previous work of Tzahor et al. is not feasible as in their study, Tzahor et al. performed the multi-class classification using full length psbA sequences, while tests with short fragments (100, 200 and 300 bp) were limited to binary classification only. For phoH, the mean precision resulted to be equal to 0.98, mean recall 0.98 and mean \( F_1 \) 0.98. No optimal probability estimate cut-off was applied for phoH sequences while the parameter cost was set to 8.

3.6 Classification of Sequences Extracted from the Metatranscriptome

In order to extract psbA and phoH fragments from the metatranscriptome data of Ottesen et al., we firstly...
Figure 2: Example of Empirical Mode Decomposition of a simulated signal. Each IMF represents a different oscillatory component/frequency of the original signal and can be used to identify the different underlying processes that are responsible for generating the original signal. Oscillatory components are extracted in decreasing order of frequency.

3.7 Empirical Mode Decomposition

Empirical mode decomposition (Huang et al., 1998), is a data-driven method for analysing of non linear and non stationary time frequency data, such as natural signals.

The EMD process iteratively decomposes the original signal into a finite number of intrinsic mode functions (IMF), that is, functions with a single mode/frequency 2. Each IMF represents a different component/frequency of the original signal and can be used to identify the different underlying processes that are responsible for generating the original signal (Figure 2). The EMD procedure works as follows: first, all the time series local minima and maxima are identified. Interpolation then is applied to connect the local minima among them, generating the lower envelope of the data. The same is performed on the local maxima, generating the upper envelope. Subsequently, the mean value of the envelope is calculated and subtracted from the original signal. This procedure takes the name of sifting and produces an IMF. To be considered valid, an IMF needs to satisfy the following conditions: (i) the number of extrema and of zero-crossings must differ by no more than one (ii) both lower and upper envelopes must have a mean equal to zero. The sifting procedure is repeated until no further IMF can be extracted from the original signal or the specified terminating criterion is met. IMFs are extracted in decreasing frequencies levels. EMD is a relatively recent approach that still holds some drawbacks. For instance, EMD may be prone to suffer from sampling errors as those could lead to incorrect placement of extrema and therefore lead to inaccurate IMFs. Similarly, the usage of different interpolation methods can also lead to slight differences in the algorithm results, particularly in terms of flexibility and smoothness of the IMFs (Bagherzadeh and Sabzehparvar, 2015). IMFs are also challenging to interpret in absence of knowledge about the underlying system, and IMFs of different orders on different time series may not be capturing the same phenomena. In addition, analogous information may end up being contained in multiple IMFs and there can sometimes be lower-order IMFs that are just spurious fluctuations (which have the purpose of correct errors on other IMFs, so that they can sum up to the original signal) (Chambers, 2015). In this work EMD was run using the EMD R package (Kim and Oh, 2009), allowing a maximum number of sifting iterations equal to 50, with a periodic type of boundary and constructing the envelopes by interpolation. No meaningful changes in the results were observed when different boundaries and interpolation methods were tested.

4 RESULTS AND DISCUSSION

A total of 20,235 psbA and 5,008 phoH short sequences (length ranging from 150 to 300b) extracted from the high-resolution metatranscriptome time course (Ottesen et al., 2014) were classified using the SVM-based classifier. As expected, retrieved psbA transcripts were most abundant in samples collected around mid-day and less abundant in samples collected around mid-night. After SVMs-based classification, most of the psbA sequences were identified as Prochlorococcus, including high light and low light ecotypes (54% and 12% on average respectively and subsequently merged in one category for downstream analysis) while only a minor fraction was classified as Synechococcus psbA (1% on average). The re-
main sequences were classified as *Synechococcus* virus (27% on average) and *Prochlorococcus* virus (6% on average). Given that *Synechococcus* represents only a small fraction of the bacterioplankton in the time series of the North Pacific Subtropical Gyre (Ottesen et al., 2014), we hypothesized that the viral *psbA* clustering into this group were not from *Synechococcus* viruses but instead from viruses that infect both *Prochlorococcus* and *Synechococcus*. Indeed, some cyanophages have shown to have a broad host range and infect strain from both genera (Sullivan et al., 2003). For example, the *psbA* found in the cyanophage P-SSM1, which was isolated using the *Prochlorococcus* strain MIT 9313 clustered with the *Synechococcus* virus group (Sullivan et al., 2006). Consequently, the two virus groups were redefined as Virus Group I (VG1; *Synechococcus* virus) and Virus Group II (VG2; *Prochlorococcus* virus). *Prochlorococcus* and *Synechococcus* are referred to as Pro Bac and Syn Bac.

*PhoH* sequences retrieved from the metatranscriptome did not show a change in abundance based on the time of the day (data not shown). The lack of variation for the *phoH* gene is probably due to the fact that light has a limited influence on the phosphate uptake. After SVMs-based classification, most *phoH* transcripts were assigned to heterotrophic phages (HP, 31% on average) and autotrophic bacteria host (A Host, 31% on average), followed by heterotrophic bacterial host (H Host, 21% on average) and autotrophic phages (AP, 14% on average). Few *phoH* sequences were classified as eukaryotic phytoplankton viruses (2% on average) in agreement with the notion that most of the primary production in the North Pacific Subtropical was supported by cyanobacteria. As a consequence, the *phoH* group including viruses infecting eukaryotic phytoplankton was removed from downstream analysis.

After classifying the sequences into their respective subgroups, we performed an empirical mode decomposition (EMD) (see Methods) on the transcriptional profiles of the subgroups, from which different diel patterns emerge within *psbA* and *phoH* transcripts. EMD decomposed each profile of classified AMGs into simpler harmonic waveforms or intrinsic mode functions (IMFs). While the EMD method lacks a formal procedure to associate a corresponding meaning/phenomena to each IMF, complicating their interpretation in situations of total absence of knowledge about the system, in this case the underlying driving forces were not completely unknown. We identified a plausible biological explanation to each IMF to support our findings (in Figure 3 the complete EMD results for VG2 is shown as example). We hypothesize that the 1st order of IMFs captured most of the stochastic noise in each time series. The IMFs of the 3rd order could clearly identify a diel pattern in the expression of all AMGs (Figure 4), while the 2nd order IMF is able to detect what we believe are differences in the population heterogeneity among different groups (Figure 5).

4.1 Diel Pattern Difference between *Synechococcus* and *Prochlorococcus*

The IMFs of the 3rd order identified a diel pattern in the expression of all AMGs (Figure 4). The peak of expression of *psbA* from *Prochlorococcus* consistently occurred in the morning while a less-pronounced cycling of *Synechococcus* *psbA* transcripts occurred later in the afternoon (Figure 4A). This is consistent with a published dataset obtained using real-time PCR (Mella-Flores et al., 2012). The remarkable difference between the *psbA* expression levels of *Synechococcus* and *Prochlorococcus* is probably caused by the fact that *Prochlorococcus* is not able to withstand solar irradiance as high as *Synechococcus*. Consequently, as *Prochlorococcus* D1 protein is more sensitive to light, *psbA* needs to be highly expressed during the time of the day with the more irradiance (Mella-Flores et al., 2012).

Both VG1 and VG2 peaked during the daytime but the peak for VG2 was precisely matched to the peak in *Prochlorococcus* *psbA*. On the other hand, maximal expression of VG1 was more variable, possibly as a result of the variable proportions of the infected strains of *Prochlorococcus* and *Synechococcus* hosts. While transcripts from *psbA*-carrying virus had a very tight coupling to the time of the day, much less diel
effect was observed in *phoH* transcripts (Figure 4B). VG1 and VG2 peaked at similar times as their hosts. On the contrary, the HP transcripts consistently peaked later than the hosts homologs and, in general, the cycling of the *phoH* transcripts was less strong in tune with the daily light cycle.

### 4.2 The Population Structure Shows Different Viral Groups

Underlying the primary diel harmonics of the 3rd order IMF, the 2nd order was able to detect differences in the population heterogeneity among different groups. This was particularly evident during the final 30 hours of the time series, when a significant increase in temperature and salinity had been previously associated to a change in the transcriptional profile of SAR324 (Ottesen et al., 2014). Here, the shift in environmental conditions is detected as increased frequency in IMF2 for both *Prochlorococcus* and *Synechococcus* (Figure 5A), suggesting that the population heterogeneity of the 2 groups has expanded to include more ecotypes.

Differences in population heterogeneity are also observed as differences in IMF2 frequencies in VG1, VG2, AP and HP (Figure 5B). Interestingly, the highest population heterogeneity (i.e. the highest IMF2 frequency) was observed in VG2, which also had the strongest correlation between AMG peak and the time of the day. Taken together, these observations suggest that VG2 includes many viral quasispecies with very narrow host ranges and a very precise control replication timing as a function of the time of the day and the low stability of the *Prochlorococcus*-type D1 protein. VG1 was less diverse but also appeared to have broader host range and the expression peak shifted in response to increased host diversity towards the end of the time-series.

The timing of the peak expression of the viral *psbA* also provided insights into the timing of cell lysis. Based on studies conducted in the laboratory, the viral *psbA* is transcribed at the end of the lytic cycle (Clokie et al., 2006b) and if replication of viral DNA is delayed until light, viral *psbA* expression should be minimal in the morning and slowly increase throughout the day.

Therefore, notwithstanding the differences between VG1 and VG2, our findings are consistent with the hypothesis that most of the cell lysis and viral shedding occurs in the end of the day. Using SeaFlow cytometry which gives real-time continuous observations of cells abundance, Ribalet et al. (Ribalet et al., 2015) showed that *Prochlorococcus* cell numbers were higher in the day and sharply decrease at night, suggesting predation (viral and grazing) as the cause of the oscillation. Our data further support their...
hypothesis that viral infection might be a player in the synchronized oscillations of Prochlorococcus abundance from surface water of a Pacific Gyre (Ribalet et al., 2015).

5 CONCLUSIONS

Using a combination of SVMs based classification of short DNA sequences and EMD, we identified the existence of light-dark oscillations in the viral infection of cyanobacterial populations in the North Pacific Subtropical Gyre which can affect Prochlorococcus cell numbers and activity from surface water. Ottesen et al. (Ottesen et al., 2014) have observed diel cycling in the expression of genes from heterotrophic bacterioplankton and suggested that factors other than light might be linking the diel behaviours of autotrophs and heterotrophs.

Here we expand on that study to include cycling of expression of the viral assemblages and, based on the expression patterns of psbA and phoH, are able to identify major viral groups that differed in their response to light-dark cycles, population structure and their host range. Undoubtedly many other groups of cyanophages exist which do not carry those AMGs and which might or might not display circadian cycling, but a previous study suggests that as much as 88% of cyanophages do indeed carry a copy of the psbA gene (Sullivan et al., 2006). Because cyanophage ecotypes carrying similar AMGs are likely to occupy very similar ecological niches, it is conceivable that the existence of different replication patterns might allow coexistence of multiple ecotypes by suppressing competitive exclusion. We posit that cyanophages of the VG1-type overcome competitive exclusion with a broad host range and be able to initiate the replication cycle at different times of the day. On the other hand, the tight coupling between replication and lower stability of the Prochlorococcus-like D1 protein requires VG2 to have a much larger genetic diversity and a higher degree of specialization in host targets.

In addition to the ecological insights that this approach has provided in understanding cyanophage populations and their hosts, similar classification and decomposition analyses may be used to identify fundamental frequencies of natural processes from other time series data that otherwise would be overlooked.

ACKNOWLEDGEMENTS

The authors would like to acknowledge financial support from Singapore Ministry of Education Academic Research Fund Tier 3 under the research grant MOE2013-T3-1-013, Singapore National Research Foundation under its Marine Science Research and Development Programme (Award No. MSRDP-P13) and the Singapore Centre on Environmental Life Sciences Engineering (SCLESE), whose research is supported by the National Research Foundation Singapore, Ministry of Education, Nanyang Technological University and National University of Singapore, under its Research Centre of Excellence Program. The authors would like to thank Fabio Stella, Rohan Williams and James Houghton for their valuable feedbacks.

REFERENCES


Supervised Classification of Metatranscriptomic Reads Reveals the Existence of Light-dark Oscillations During Infection of Phytoplankton by Viruses


