Spatial Analysis and Visualization of Genetic Biodiversity

Robert G. Beiko[1,*], Jacqueline Whalley[2], Suwen Wang[1], Harman Clair[1], Greg Smolyn[1], Sylvia Churcher[1], Michael Porter[1], Christian Blouin[1], and Stephen Brooks[1]

[1]Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia, Canada
[2]School of Computer Science, Auckland University of Technology, Auckland, New Zealand

[*]Contact: Faculty of Computer Science, Dalhousie University
6050 University Avenue, Halifax, Nova Scotia, Canada B3H 1W5
Tel: +1 902 494 8043 Email: beiko@cs.dal.ca

Bacteria, viruses, and single-celled eukaryotes are vital components of the Earth's biodiversity. Until recently, it has been nearly impossible to study these microorganisms outside of a laboratory setting due to their small size, immense population numbers, and enormous diversity. However, recent advances in DNA sequencing technology are opening up the microbial world for fine-scale examination. New ways of assessing microbial biodiversity and ecology have already produced stunning results: species counts in some habitats are thought to number in the tens of thousands, and microbial evolution appears to be driven in part by widespread sharing of genes among organisms. By comparing genetic data collected from many sample sites, we can investigate the interactions between the environment and microbial diversity, including the possible effects of spatial correlations among sites. Our ability to understand and learn from these data sets is limited by the availability of tools for analysis and visualization of sample sites, the distribution of microbial species and gene functions, and the evolutionary context of these organisms.

We have developed GenGIS, a free and open source GIS package that provides tools for spatial visualization and analysis of microbial genome data. GenGIS renders maps in a three-dimensional OpenGL environment, and makes extensive use of other open-source software as well as publicly available geographic and genetic data. Here we describe the features of the first release version of our GIS, and use it to test questions about microbial biodiversity in the world's oceans.

Introduction

Microbes have been found in almost every environment on Earth. In addition to their familiar roles in disease and industrial processes, microorganisms are vital to the processes that sustain life, including energy capture from light and the breakdown and synthesis of essential carbon- and nitrogen-containing molecules. Microorganisms can survive at extremes of temperature, pressure, and salinity, and are present in polluted environments that contain hazardous materials such as heavy metals and polychlorinated biphenyls and ethenes (PCBs and PCEs). In some cases, these organisms can break down or neutralize these contaminants: for instance, the bacterium *Dehalococcoides ethenogenes* obtains energy from the breakdown of PCE compounds, reducing them to the harmless compound ethane (de Bruin et al, 1992). The community of microorganisms at a given site can also serve as a sensitive indicator of environmental change, as microbes that predominate at a pristine habitat may be outcompeted by new species if that site is subjected to pollution or other changes. In a survey of the central Pacific Line Islands, coral reef-associated microbial communities were shown to vary dramatically with differences in coral reef health, with potentially pathogenic species dominating where human impacts were highest (Sandin et al, 2008).

The utility of microorganisms in environmental assessment and bioremediation has been demonstrated in many studies, but the assessment and discovery of interesting microbes has been limited by their small physical size, immense numbers, and lack of distinguishing features. The advent of cheap and reliable DNA sequencing technology in the last 15 years has transformed the study of these organisms; whereas before an organism's potential had to be assessed through painstaking laboratory work, we can now infer a great deal of function from an organism by sequencing its genes and comparing them to genes whose function has already been experimentally determined. Moreover, while traditional techniques for microbial community assessment require expensive experiments and microscopy, a 'snapshot' of the species present can now be rapidly and economically acquired by sequencing genes directly from an environmental sample. Certain genes can be used as markers of diversity, if they satisfy the following three criteria: (1) they must be present in all groups of interest (e.g., species); (2) they must evolve sufficiently quickly that their DNA sequences can distinguish among groups; and (3) they must not evolve too quickly, or they would not be recognizable as 'the same gene' in distantly related species. Most known genes do not satisfy these criteria, either because they are not present in all organisms or because they evolve too quickly to be of use as a diagnostic. But certain fundamental genes can be used, including components of the globally conserved protein synthesis machinery, and genes involved in energy production: these are known as 'marker' or 'barcode' genes.

These sequencing techniques are now being exploited to generate massive genetic data sets representing the entire genetic repertoire of a given organism (e.g. the complete genome of *Escherichia coli*), or a snapshot of microbial species composition and function directly from environmental samples. The most notable example of the latter to date is the Global Ocean Sampling (GOS) expedition, which was carried out from 2003-2006 and collected and sequenced more than 120 environmental samples during a circumnavigation of the Earth. This massive dataset roughly doubled the known repertoire of functional genes (Rusch et al, 2007), but building a biochemical and ecological understanding of these data sets requires new approaches that have not yet been extensively applied in microbiology and bioinformatics. As we aim to understand this and other pioneering datasets others are emerging at an accelerating pace, offering great opportunity but equally great technical challenges.

The geographic nature of these new 'metagenomic' studies offers the possibility of discerning what environmental and spatial factors drive the composition and function of microbial communities. To this end we have developed GenGIS, a new open source geospatial information system that combines three-dimensional map visualizations with a statistical language for data analysis. Although potentially useful in a range of applications, GenGIS is focused on the analysis of genetic data collected from the environment, the construction and testing of hypotheses, and the visualization of results.

GenGIS implementation and features

Two existing Web applications are available for visualization and retrieval of DNA sequence data. Micro-Mar (Pushker et al, 2005) and megx.net (Lombardot et al, 2006) both offer a two-dimensional map interface with geographic references for sequenced organisms and environmental genomic samples. Additional features of megx.net include MetaLook, a three-dimensional environment for concurrent visualization of habitat and genetic data; and Geographic BLAST, which allows a user to search for genes of particular function within a restricted geographic area. An attractive feature of megx.net is its integration of environmental data from the U.S. National Ocean Atlas, which associates genetic samples with relevant temperature, nutrient concentration, and other parameters. Digital map data in Micro-Mar are derived from Google Maps, while megx.net is again based on the National Ocean Atlas.

Our aim in developing GenGIS is to provide a richer, three-dimensional environment that supports data visualization, statistical analysis, and integrated support for other bioinformatics applications. To support the analysis of large data sets and a richer range of 3D features, GenGIS has been developed as a standalone application that makes use of a user's local computing power. This arrangement also simplifies the analysis of private (pre-publication or commercially sensitive) data, since there is no need to upload these data to a central server.

GenGIS has a model-view-controller architecture, with key data model elements (digital elevation, geographic locations and environmental properties of samples, and the properties of sequences collected from each site) implemented as C++ objects. Boost <http://www.boost.org> shared pointers are used extensively to connect different model elements without duplicating storage, and eliminating the need for reference tracking. Many user events can be initiated either from the console, through the menu system, or through direct interaction with the OpenGL canvas (see below); these requests are translated into a consistent API function call by the controller, which then accesses the model data. This approach simplifies the addition of new features, since new methods to interact with data can be implemented without needing to understand or modify the data model.

Map, sample, and sequence data sets

Three types of input data are required by GenGIS: digital elevation data, information about the location of samples, and information about the sequences collected at each site. GenGIS uses the Geospatial Data Abstraction libraries (GDAL): <http://www.gdal.org> to open and convert digital elevation files, allowing the user to load and view elevation data in many formats, including those used to distribute the global GTOPO30 and SRTM (Farr et al, 2007) data sets. The location and sequence files consist of comma delimited text fields, where the first field in each case is a unique identifier. The second and third entries in the location file give the coordinates of the sample. Locations are expressed either as decimal degrees of latitude and longitude, or as northings and eastings in the UTM coordinate system. The remaining fields in each file can contain arbitrary information as specified by the user; these columns can be used as limiting criteria for the generation of subsets based on sample and sequence properties and as contrasting categories for statistical analysis.

Environment and user interface

The core visualization functions of GenGIS were implemented in C++/OpenGL, with digital elevation data visualized in three dimensions. The graphical user interface uses the wxPython toolkit which provides standard window manipulation tools, a menu bar, and a set of dockable panels. The environment comprises several panels:

- The **OpenGL canvas** is the main 3D visualization environment, which displays the digital map, clickable navigation controls, and any graphical representations of sample and sequence data. Efficient terrain rendering is achieved by reducing the level of display detail as distance from the camera increases (Wang et al, 2007).

- A set of **legend windows** can be used to display and change the colours that are associated with map altitude and categories of sample and sequence data.
- A **console window** allows the user to enter commands from the keyboard and offers standard Python functions, custom API commands that have been developed for GenGIS, and statistical analysis functions that are made available through the Rpy libraries <http://rpy.sourceforge.net>. The user can execute custom scripts directly in the console, allowing similar analyses to be repeated on different data sets.

The user can navigate (pan, zoom, tilt, rotate, and incline) in the 3D environment through the menu bar, by interacting with the navigation controls in the OpenGL canvas, using hotkeys that are mapped to menu functions, or by entering API functions directly into the Python console. The last of these options allows the user to save and run scripts that automate the loading and manipulation of data.

Data analysis

An important component of GenGIS is the integration of the open-source statistical language R into the Python console via the Rpy libraries. Users of GenGIS can use any statistical function that is available in R or the many specialized libraries available through CRAN <http://cran.r-project.org> to analyze their data. Three console API commands allow a user to create R plots and display them directly in the OpenGL canvas. The user can generate a plot that summarizes the sequence data from a single sample, a series of plots that provide the same type of summary for each sample, or a single plot that compares a property from each sample. These plots, and any data visualizations generated natively by GenGIS, can be automatically laid out within the OpenGL canvas or can be manipulated by the user to achieve the desired positioning of elements.

Application to environmental genomic data

Standards concerning public access to genetic data were laid down in the early to mid 1980s. With very few exceptions, any publication that describes the sequencing and analysis of DNA or the corresponding protein sequences must be accompanied by a deposition of these data in a public repository. The principal repository for such data is the U.S. National Center for Biotechnology Information <http://www.ncbi.nlm.nih.gov>, but new resources are emerging to house different types of data. Complementary to NCBI are resources such as the CAMERA project (Seshadri et al, 2007), which serves marine genomic and metagenomic data. Open data policies have spurred the development of bioinformatics and allied fields, and have been of central importance to our emerging understanding of biodiversity at the genetic level. In this section we demonstrate the application of GenGIS to the Global Ocean Sampling data set.

Visualizing genetic diversity

The GOS project was unprecedented in its scope, collecting hundreds of millions of short DNA sequence ‘reads’ from over 100 sites in the world’s oceans. In addition to profiling species diversity in different habitats, GOS researchers were interested in genes encoding proteins with particularly important functions. One question of particular interest was the prevalence of mechanisms to harvest energy from light, because energy production is of fundamental importance to all life, and because of the potential for light-harvesting proteins to be commercially exploited as a source of clean energy (Hampp et al, 2004). The proteorhodopsin family of proteins, although discovered only in the last decade, play a significant role in harvesting light energy in the world’s oceans.

DNA sequences identified as encoding proteorhodopsin proteins were recovered from many samples, but given the wide diversity of bacteria that possess these proteins, it is useful to ask whether the same organisms play the same vital ecological role in every sampled habitat, or whether different species dominate under different environmental conditions. We used GenGIS to examine this question in a targeted set of samples collected near the Galapagos Islands, in the territorial waters of Ecuador. Eleven samples comprising several types of coastal habitat were collected near several islands in the archipelago, and predicted protein sequences from these sites were compared to known sequences from the NCBI database using the BLAST algorithm (Altschul et al, 1997). Since the protein sequences from NCBI were all isolated from known species, the GOS environmental sequences could be assigned the same classification as the reference sequence with which they shared the highest degree of similarity.

Figure 1 shows a top-down view of the eleven Galapagos Island GOS sample sites, overlaid on SRTM digital elevation data. The breakdown of sequences by high-level grouping (phylum, a classification also used to distinguish groups of animals such as arthropods and vertebrates) shows that most sites are dominated by phylum Proteobacteria, a grouping that includes many known pathogens as well as environmentally important organisms. The exception to this pattern is one site near Floreana Island in the south, which is dominated by phylum Actinobacteria.

When these samples are examined at a more-precise level of classification, the majority of Proteobacterial sequences are shown to belong to genus *Pelagibacter,* a group that is known to be widespread in the world's oceans. But the Floreana Island sample is dominated by sequences that cannot be assigned a more-precise classification due to their limited similarity to other sequences in the database. These sequences are consequently assigned to "Unclassified Actinobacteria" and their species of origin remains unknown.

What is the cause of the drastic difference in biodiversity and dominant phylum at the Floreana site? Environmental data collected alongside the sequence samples show that this site has a much higher temperature (37.6 degrees C) relative to the other samples, none of which has a temperature greater than 28 degrees C. This sample was also collected from a lagoon where salt concentrations are nearly double those of the other coastal sites. Organisms that live in this environment must therefore be adapted to higher temperature and salinity: *Pelagibacter* cannot thrive in this environment, but the unclassified Actinobacterial species can.

Spatial analysis of diversity

Another important question in microbial ecology is the spatial scaling of microbial diversity; in other words, at what distances do we expect to see spatial correlations in the relative abundance of different types of microbes? This question is perhaps best captured by the hypothesis articulated by Bass Becking (1934), which is most accurately translated as "Everything is everywhere, but, the environment selects". For this hypothesis to be true, microbes must be able to transmit themselves globally, and survive in those habitats for which they are adapted. The alternative is that at least some microbes are restricted in their geographic range, and that the same ecological 'job' would be carried out by different species in different parts of the world. Marker gene studies have often suggested that some species are cosmopolitan, i.e. the exact same DNA sequence has been found in many instances of the same type of habitat, spread over the entire globe (Hedlund and Staley, 2003). However, other studies have revealed that some hot spring (Papke et al, 2003) and soil

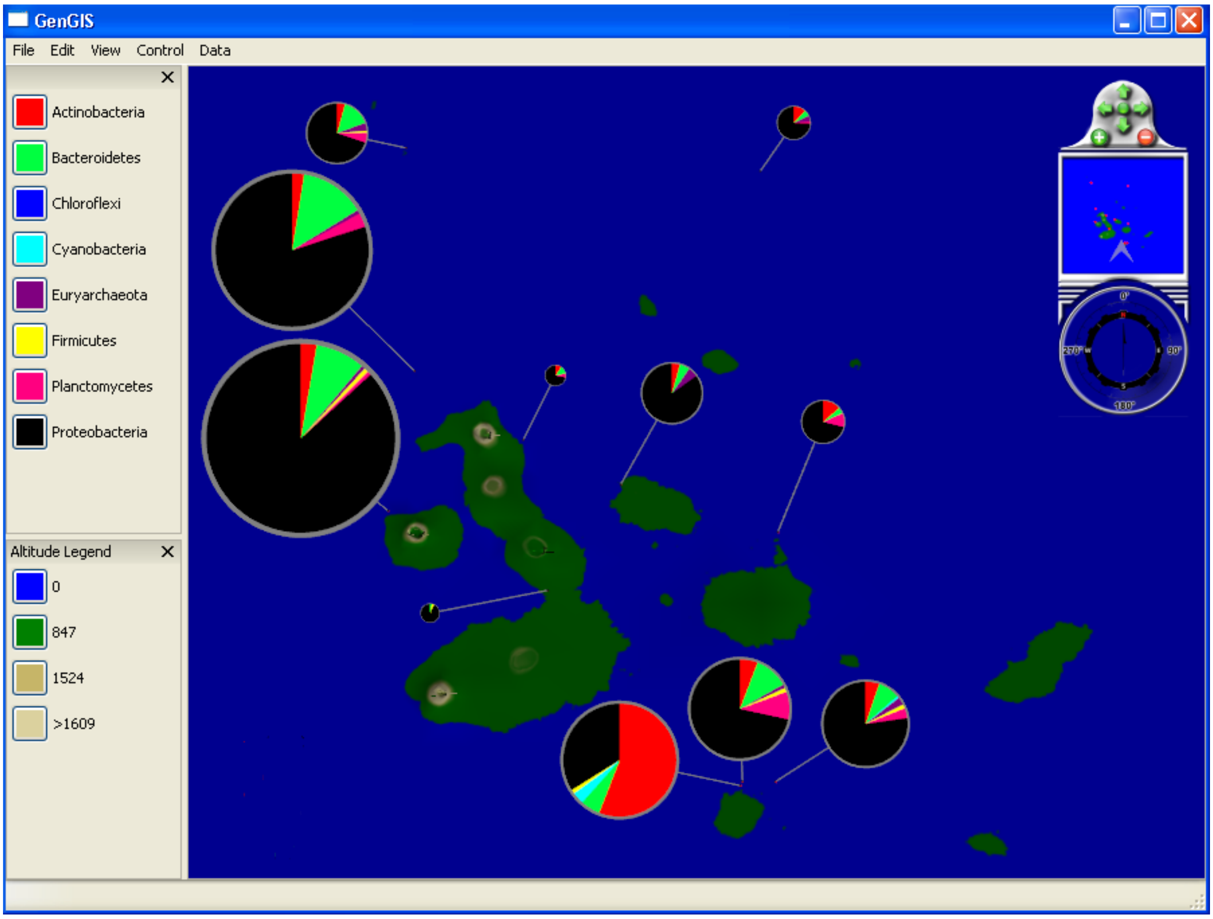

Figure 1: phylum association of proteorhodopsin-like sequences collected from eleven sample sites in the Galapagos Islands archipelago. The main frame contains the OpenGL canvas and displays digital elevation data, pie charts showing the total count (size) and relative proportion (wedge size) of proteorhodopsin sequences, and control elements including the minimap. Classification and elevation legends are shown in the left-hand panels.

(Cho and Tiedje, 2000) bacteria have limited distributions. It may be that the spatial relationships among bacteria are dependent on their mode of dispersal (Ramette and Tiedje, 2007).

We used GenGIS to test for spatial relationships in a subset of GOS samples collected along the eastern coast of North America, stretching from Nova Scotia south to the Caribbean Sea (Figure 2). Species presence and absence was characterized using 16S ribosomal DNA genes as markers, and classifications were assigned in a manner similar to that used for proteorhodopsin sequences above. In this case we examined the distribution of bacterial groups that were frequently observed (>5% of all samples) in the Atlantic transect data; the groups examined comprised several common genera including *Prochlorococcus* and *Pelagibacter*, the phyla Proteobacteria and Firmicutes, and an ‘other’ group that contained all remaining sequences. The Bray-Curtis index (BCI: Bray and Curtis, 1957) was used to assess the similarity of species compositions between pairs of sites; the BCI can range from 0 (the samples contain completely different groups) to 100 (both samples contain the same groups in exactly identical proportions). We performed a linear regression analysis of BCI versus geographic distance for each pair of samples (i.e., a Mantel test: see e.g. Horner-Devine et al, 2004). A statistically significant relationship would suggest a relatively large range (> 500 km)

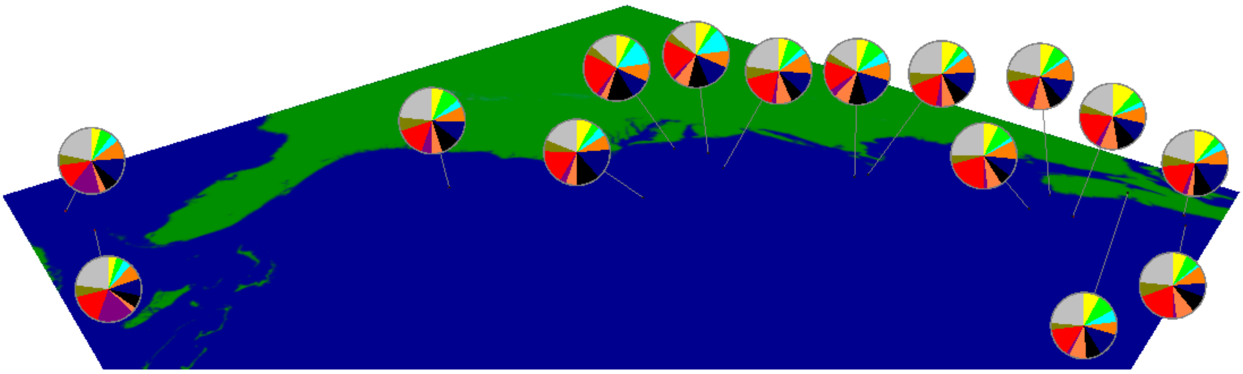

Figure 2: relative frequency of 16S ribosomal DNA marker gene sequences at 15 different GOS sample sites taken along the east coast of North America.

over which microbial populations are correlated with one another. However, no statistically significant relationship was observed (Mantel p = 0.25), indicating a lack of correlation among relatively close sites. If spatial correlations exist for the groupings examined here, they are likely to be observable only at distances much smaller than 500 km.

Conclusions

We have prototyped the data integration and analysis capacity of GenGIS using the Global Ocean Sampling dataset described by Rusch et al (2007). Visual analysis of the distribution of proteorhodopsin sequences suggested a strong influence of the environment on the dominant sequence type: this observation suggests a hypothesis that could be tested with more-intensive sampling of these types of habitat. Such observations have important implications for ecological stability; if the vital process of light energy capture is carried out by different organisms depending on the type of habitat, then changes to these habitats (e.g. through climate change or pollution) could have very different effects on the community of organisms that reside there. The discovery of organisms and sequences is also of interest in bioengineering; proteins that can tolerate extremes of temperature or salinity will be better suited to use in certain types of industrial process. Although there was considerable variation in biodiversity along the Atlantic seaboard, our spatial analysis of bacterial diversity did not reveal any correlation between nearby sites. Sampling at smaller spatial scales will be necessary to further test for spatial correlation among populations, and spatial effects will need to be examined in concert with other factors such as latitude and salinity.

Our long-term goals for the development of GenGIS include closer integration of sequence analysis tools, including phylogenetic methods that reconstruct the evolutionary relationships among genes and genomes. When combined with indices of space and time, such methods will allow users to (for instance) track mutations that occur during an outbreak of disease, and infer the ways in which microbial communities have evolved to exploit new habitats. The public release of GenGIS is scheduled for September 2008, and will include full open source and compiled versions for Windows and Mac OS X.

Acknowledgments

The authors would like to thank Olga Zhaxybayeva, Adrian Sharma and Jeremy Koenig for organizing the Global Ocean Sampling data sets, and Hong Gu for statistical advice. This work was supported by the Natural Science and Engineering Research Council of Canada, Genome Atlantic, and the Auckland University of Technology.

References

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ 1997, ‘Gapped BLAST and PSI-BLAST: a new generation of protein database search programs’, *Nucleic Acids Research*, vol. 25, no. 17, 3389-402.

Baas Becking, LGM 1934, *Geobiologie of inleiding tot de milieukunde*, W.P. Van Stockum & Zoon, The Hague, the Netherlands.

Bray R, Curtis T 1957, ‘An ordination of the upland forest communities of Southern Wisconsin’, *Ecological Monographs*, vol. 27, 325-49.

de Bruin WP, Kotterman MJJ, Posthumus MA, Schraa G, Zehnder AJB 1992, ‘Complete biological reductive transformation of tetrachloroethylene to ethane’, *Applied Environmental Microbiology*, vol. 58, no. 6, 1996–2000.

Cho JC, Tiedje JM 2000, ‘Biogeography and degree of endemicity of fluorescent *Pseudomonas* strains in soil’. *Applied Environmental Microbiology*, vol. 66, no. 12, 5448-56.

Farr TG, Rosen PA, Caro E, Crippen R, Duren R, Hensley S, Kobrick M, Paller M, Rodriguez E, Roth L, Seal D, Shaffer S, Shimada J, Umland J, Werner M, Oskin M, Burbank D, Alsdorf D 2007, ‘The Shuttle Radar Topography Mission’, *Reviews of Geophysics*, vol. 45, RG2004.

Lombardot T, Kottmann R, Pfeffer H, Richter M, Teeling H, Quast C, Glöckner FO 2006, ‘Megx.net--database resources for marine ecological genomics’, *Nucleic Acids Research*, vol. 34, D390-3.

Hampp N,Oesterhelt D 2004, ‘Bacteriorhodopsin and its potential in technical applications’, in C Niemeyer and C Mirkin (eds.), *Nano-bio-technology : Concepts, Applications and Perspectives*, Weinheim:Wiley-VCH-Verlag, New York.

Hedlund BP, Staley JT 2003, ‘Microbial endemism and biogeography’, in AT Bull (ed.) *Microbial Diversity and Bioprospecting*, ASM Press, Washington, DC.

Horner-Devine MC, Lage M, Hughes JB, Bohannan BJ 2004, ‘A taxa-area relationship for bacteria’, *Nature*, vol. 432, no. 7018, 750-3.

Papke RT, Ramsing NB, Bateson MM, Ward DM 2003, ‘Geographical isolation in hot spring cyanobacteria’, *Environmental Microbiology*, vol. 5, no. 8, 650-9.

Pushker R, D'Auria G, Alba-Casado JC, Rodríguez-Valera F 2005, ‘Micro-Mar: a database for dynamic representation of marine microbial biodiversity’, *BMC Bioinformatics*, vol. 6, 222.

Ramette A, Tiedje JM 2007, ‘Biogeography: an emerging cornerstone for understanding prokaryotic diversity, ecology, and evolution’, *Microbial Ecology*, vol. 53, no. 2, 197-207.

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K, Beeson K, Tran B, Smith H, Baden-Tillson H, Stewart C, Thorpe J, Freeman J, Andrews-Pfannkoch C, Venter JE, Li K, Kravitz S, Heidelberg JF, Utterback T, Rogers YH, Falcón LI, Souza V, Bonilla-Rosso G, Eguiarte LE, Karl DM, Sathyendranath S, Platt T, Bermingham E, Gallardo V, Tamayo-Castillo G, Ferrari MR, Strausberg RL, Nealson K, Friedman R, Frazier M, Venter JC 2007, ‘The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific’, *PLoS Biology*, vol. 5, no. 3, e77.

Sandin SA, Smith JE, Demartini EE, Dinsdale EA, Donner SD, Friedlander AM, Konotchick T, Malay M, Maragos JE, Obura D, Pantos O, Paulay G, Richie M, Rohwer F, Schroeder RE, Walsh S, Jackson JB, Knowlton N, Sala E 2008, ‘Baselines and degradation of coral reefs in the Northern Line Islands’, *PLoS ONE*, vol. 3, no. 2, e1548.

Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M 2007, ‘CAMERA: a community resource for metagenomics’, *PLoS Biology*, vol. 5, no. 3, e75.

Wang S, Beiko RG, Brooks S 2007, ‘Collapsible 3D Terrains for GIS Visualization’, in *Geovisualization 2007*, NUI Maynooth.