11-30-2005

Collaborative Research: Toward Environmental Genomics: Can We Estimate Bacterial Diversity in the Ocean?

Daniel L. Distel
Principal Investigator; University of Maine, Orono

Final Report for Period: 09/2002 - 08/2006 **Submitted on:** 11/30/2005
Award ID: 0221224
Principal Investigator: Distel, Daniel L.
Organization: University of Maine
Title:
Collaborative Research: Toward Environmental Genomics: Can We Estimate Bacterial Diversity in the Ocean?

Project Participants

Senior Personnel

Name: Distel, Daniel
Worked for more than 160 Hours: Yes
Contribution to Project:

Post-doc

Graduate Student

Name: Luyten, Yvette
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Nair, Nitin
Worked for more than 160 Hours: No
Contribution to Project:

Undergraduate Student

Technician, Programmer

Other Participant

Research Experience for Undergraduates

Organizational Partners

Other Collaborators or Contacts

Activities and Findings

Research and Education Activities:
Goals and objectives
Our proposal was structured around 3 basic questions pertaining to microbial community structure with several attached objectives:

1) What is the diversity of co-occurring ribotypes in a natural bacterioplankton community?
The objective was to compare sequence diversity in 16S and 23S rDNA clone libraries obtained from the same coastal bacterioplankton sample in order to compare diversity estimates. We also planned on using modified PCR techniques to constrain generation of artificial sequence types

by PCR.

2) Can we sample genomic diversity within individual ribotypes and between closely related ribotypes in natural environments?
The objective was to identify what genetic, ecological or phylogenetic unit is actually represented by unique or closely related rRNA sequences. The plan was to use our newly developed capture and walk technique that allows the use of oligonucleotides specific for a particular ribotype to pull large genome fragments (> 10 Kb) from environmental samples. These can be sequenced and specific probes complementary to their ends can be designed and employed to capture contiguous sequence pieces. Thus, clone libraries that are samples of the co-existing diversity within identical and similar ribotypes can be assembled and the diversity of associated genes explored without prior cultivation of the organisms.

3) What is the extent of genomic diversity within identical and similar ribotypes in natural environments and what mechanisms contribute to its generation?
The objective was to explore the types of sequence variation that dominate the genomes of co-existing ribotypes in natural communities using sequence analysis of unique ribotype lineages captured and assembled from the environment.

Experiments to address questions
Question 1
1) To address question 1, we carried out extensive sequencing of a 16S rRNA gene library constructed from a coastal environment (our model site at the Plum Island NSF LTER centered at the MBL). Two clone libraries were constructed from the same sample and over 2,000 clones were sequenced. The two libraries differed in how the genes were amplified prior to cloning to estimate the contribution of PCR induced errors to sequence diversity estimates. Thus, one library was amplified under æstandard conditionsÆ, i.e., conditions usually used for amplification of environmental DNA. The second library was constructed with the goal to minimize chimera and heteroduplex formation with protocols that we have previously optimized or developed.
2) We also developed a new web-based software to identify chimeric clones in well-sampled clone libraries and applied the software to the library. Furthermore, the contribution of Taq and sequencing errors were estimated by comparing all sequences to secondary structure models with a simple set of rules that allows identification of putative errors.
3) To arrive at better bounds of diversity estimates by ribotypes we conducted a detailed exploration of contribution of operon differences to overall microbial diversity estimates by ribotype determination. We aligned rRNA operons from all published bacterial and archaeal genomes with multiple operons and determined number of operons, range of nucleotide divergence and redundancy of identical sequences. These activities allowed us to estimate (i) the contribution of methodological artifacts to diversity estimation of sequences in the environment, and (ii) the overall diversity of co-existing ribotypes and their likely correspondence to co-existing genomes. As detailed below, the estimation of diversity combined with estimation of errors revealed important structural features of the community.
4) We constructed and analyzed a 23S rRNA library to compare diversity estimates with the 16S rRNA library. This made necessary the re-evaluation and design of æuniversalÆ bacterial 23S rRNA amplification primers. To this end, we compiled the most extensive alignment of 23S rRNA sequences from all major databases. We identified an optimized primer pair and used it for clone library construction. The 23S rRNA clone library was subsequently compared with the 16S rRNA libraries using several different statistical tools.

Question 2
1) Although the genome fragment capture method showed promising results we abandoned the approach in favor of a more simple approach, which allowed us to address the question more efficiently. Under the auspices of a Ph.D. thesis, a large collection of vibrio strains was isolated every month over an entire seasonal cycle. When analyzed by 16S rRNA sequencing we noted that the strains fall into æmicrodiverse clustersÆ as predicted by the results of Question 1. We thus decided to focus on one taxon (V. splendidus) for which a large number of strains (322) with identical or nearly identical (<1% difference) rRNA sequences were available.
2) We quantified by QPCR the V. splendidus population in the same samples the strains were isolated from to yield dynamics over an entire year.
3) We sequenced the Hsp60 gene for all 322 V. splendidus strains to determine allelic diversity of this housekeeping gene. The data also served to analyze population structure over the spatio-temporal scales sampled by FST and AMOVA.
4) Allelic diversity (albeit on a smaller subset of strains) was also determined and compared for the ToxR (signal transduction), ChiA (chitin degradation) genes.
5) Overall genome diversity for all 322 isolates was analyzed by pulsed field gel electrophoresis (PFGE). For a subset of 12 strains, genome size was determined.

Question 3
1) We analyzed the genes obtained for evidence of selection. The effort to identify homologous recombination rates is ongoing by a multilocus sequence typing (MLST) approach we have carried out on all our strains. This entailed sequencing of 6 housekeeping genes for all isolates.
2) We have also obtained two complete genome sequences (funded by the Moore foundation) and these will form the basis for analysis of gene content variation among the genomes (continuing effort).

Findings:

Major conclusions

We have developed and applied protocols for estimation and elimination of artifacts from diversity estimates. These showed that over half of the sequence diversity in clone libraries may be due to PCR errors and artifacts.

The ability to constrain PCR induced errors and artifacts led to the first well-bounded estimation of coexisting numbers of ribotypes in a single environment. Most importantly, our approach allowed us to observe fine-scale patterns in bacterial diversity and to formulate hypotheses regarding their origins. We unambiguously showed for the first time that the large majority of co-existing bacteria fall into clusters containing extremely closely related (microdiverse) taxa. Over half the diversity in the microbial population is in sequence clusters with <1% divergence. This observation provides rich substance for interpretation by ecological and evolutionary theory as it is consistent with the idea that such clusters arise by selective sweeps and that competition is too weak to purge diversity from within them. It is, in our view, a significant step towards identification of natural taxonomic units for microorganisms phylogenetically defined and grounded in ecological theory, which may ultimately turn out to be the microbial equivalent of the eukaryotic species.

Analysis of variation in one microdiverse taxon (V. splendidus) over an entire year provided insights into genomic diversity within a single population. Surprisingly, the population contained over 1,000 well differentiated (in size and gene sequences) genome types, each occurring on average at such low environmental concentration that unique traits among them must have negligible impact on overall population function. This raises questions to what extent differences in genome architecture among ocean bacterioplankton should always be judged as adaptive in a give environmental context. At the very least, our results suggest that ecological interpretation of genomic variation cannot be built on single genome sequences. We therefore believe that this information will be critical for the interpretation of variation encountered in bacterial isolates and provides an important step toward formulation of new theories of genome evolution and environmental selection.

We believe our findings will specifically impact the fields of microbiology, ecology, evolution and genomics. Microbiology has lagged behind in formulation and testing of evolutionary and ecological theories. For example, due to a lack of well-constrained diversity estimates it has not been possible to put observations on microbial community structure in the context of the rich body of literature assembled for animals and plants. We believe that our contributions will provide stimulus to compare community patterns among all forms of life. Furthermore, our data provide an important framework and context for interpretation of current effort to access microbial diversity in complex communities via environmental genomics.

Detailed results

- We have found 590 sequences in our library after subtraction of sequences, which were putative artifacts. When extrapolated to total diversity in the library by the Chao-1 estimator, a total diversity of 1,633 ribotypes was estimated to coexist. This is roughly 30% lower than the estimate arrived at before correction of putative Taq errors in the library. Furthermore, the library constructed by standard methods showed a further 30% higher diversity estimation so that we conclude that roughly 60% of diversity estimation by standard methods is due to artifacts.
- The significant reduction of artifacts in our diversity estimation allows more detailed analysis of community structure. We found that even after subtraction of Taq errors, which contributes to diversity estimation by creating large numbers of closely related sequences, about 50% of the sequences showed divergence of <1% nucleotide differences. Such small-scale divergence has usually been ignored in recent analysis of community structure due to the assumption that it is due to Taq error. Thus, we conclude that an excess of close relationships in ribotypes dominate the community structure.
- Our estimation of operon heterogeneity allows us to put further bounds on the overall diversity estimation. We found that 83 genomes with multiple operons contained 409 operons with 251 different 16S rRNA sequences (ribotypes). Thus, were these to be analyzed by standard cloning and sequencing, a roughly 3-fold overesimation of diversity would result. This overestimation would roughly be manifest in closely related sequences since operons largely differ in sequence by <1%. If genomes with a single operon are included in the analysis, a roughly 2-fold overestimation would result. Thus, we conclude that the overall ribotype diversity in the community analyzed is housed in approximately 650 to 950 genomes. Furthermore, we conclude that the observed microdiversity (<1% related sequences) is to a large extent explained by operon differences; however, the overall dominance of small-scale divergence persists in the community.
- The availability of a diverse set of 23S rRNA gene sequences enabled evaluation of the specificity of 45 previously published and 4 newly designed bacteria-specific primers. An extensive clone library constructed using an optimized primer pair resulted in similar gene richness but slightly differing coverage of some phylogenetic groups compared to a 16S rRNA gene library from the same environmental sample.
- Determination of ribotype diversity in the vibrio strains showed that close relationships dominate (1-2% nucleotide divergence) and confirms the observation in the total Bacteria clone libraries.
- Sequencing of three additional genes showed almost no redundancy in sequences in any of the three genes. Thus, microdiversity on the rRNA level translates into even higher diversity in protein coding genes. However, the sequence variation indicated neutral mutations to dominate.
- Since we have measured the upper bound of vibrio populations in the environment the strains were isolated from and all strains were isolated from <1ml water, we can estimate upper bounds of the sizes of clonal populations of these vibrios. This shows that average population size of strains with identical sequence in the Hsp60 gene is only 2-15 cells per ml and with a unique genome <1 cell per ml. This demonstrates that none of the variants clearly dominates and suggests that interactions between different clonal populations are likely confined to chance encounters and local competition.
- Preliminary phylogenetic analysis of the three protein coding genes revealed no consistent structure between the three different genes. For

example, if clustering was observed in one gene this was on average not reflected in the other genes. It has been postulated in the past that the effect of strong competition will be manifest in distinct sequence clusters, which are clearly separated from other such clusters. These arise by clonal diversification punctuated by occasional selective sweeps, which purify sequence variation at all loci even if they are selectively neutral. Because no such pattern is found in the vibrio strains we hypothesize that either recombination rates are so high that they shuffle genes relatively rapidly following selective sweeps; or, selective sweeps are extremely rare. We suggest that the latter is more likely and is due to competition being stochastically driven, locally confined and moderated by strong top-down effects. We are currently in the process of analyzing evidence of recombination in the vibrio isolates by using a multilocus sequence typing (MLST) approach.

Training and Development:

Overall, the project partially supported the efforts of 7 graduate students (6 in the Polz lab, 1 in the Distel Lab) 2 postdocs and 4 undergraduates. All graduate students and postdocs are women and two of the undergraduates were minorities.

Outreach Activities:

Results of this project have been integrated into outreach activities of the Ocean Genome Legacy including annual open house presentations to local community leaders and high school and secondary school teachers and will become part of a planned high school teacher mentoring program being conducted jointly by the Ocean Genome Legacy, New England Biolabs, and Ipswich High School.

Journal Publications

Acinas, S. G., V. Klepac-Ceraj, D. E. Hunt, C. Pharino, I. Ceraj, D. L. Distel, and M. F. Polz., "Fine-scale phylogenetic architecture of a complex bacterial community.", Nature, p. 2261, vol. 430, (2004). Published

Thompson, J. R., S. Pacocha, C. Pharino, V. Klepac-Ceraj, D. E. Hunt, J. Benoit, R. B. Sarma-Rupavtarm, D. L. Distel, and M. F. Polz., "Genotypic Diversity within a Natural Coastal Bacterioplankton Population.", Science, p. 1311, vol. 307, (2005). Published

Books or Other One-time Publications

Web/Internet Site

URL(s):

www.oglf.org

Description:

This site, which is currently under development, will host the OGL Marine Genome Resource Database, which will provide information on all specimens contained in this public access collection of marine genomic DNAs, DNA libraries, amplification products tissues and strains. The site will provide data describing collector, collections site, species ID, ecological parameters, global positioning, sample availability, terms of use,and links to research and publications related to the samples. The current project contributed a large collection of Vibrio isolates which will be included in the collection.

Other Specific Products

Product Type:

Physical collection (samples, etc.)

Product Description:

A collection of over 350 Vibrio isolates have been collected from Plum Island Sound. These isolates have been characterized by sequencing of molecular biomarkers and pulsed field gel electrophoresis and have associated ecological metadata.

Sharing Information:

These isolates will be deposited in the Ocean Genome Legacy Marine Genome Resource, a new open-access public genome resource collection maintained by the non-profit Ocean Genome Legacy Foundation, of which the PI is executive director.

Product Type:

Data or databases

Product Description:

A large collection of 16S rRNA and HSP 60 genes was generated from over 350 Vibrio strains collected in the course of this work from plum island sound. This is a unique data set in terms of its comprehensive coverage of a coastal bacterioplankton community and its accompanying quantitative data on ecological parameters and natural abundance over the course of one year.

Sharing Information:

These sequences have been submitted to genbank. The Genbank records will also contain link-outs to the OGL Marine Genome Resource Database (currently under development) that will allow Genbank users to access strains and strain information from this public access collection.

Product Type:

Software (or netware)

Product Description:

Clusterer: a software product designed to organize sequence data in progressive clusters based on sequence divergence estimates.

Sharing Information:

Manuscript in preparation

Contributions

Contributions within Discipline:

Although molecular data generally support the perception that bacterial diversity is vast, surprisingly little is known regarding the quantitative extent of this diversity, even within well-defined natural environments. This project has contributed significantly to the understanding of microbial diversity by showing the existance of a pervasive pattern of ribotype clusters in a coastal bacterioplankton community and by further exploring the internal diversity within one such cluster. These investigations allowed the significance of such clusters to be explored within the ecological context of the water column. The results reveal extraordinary diversity even within such clusters that by existing criteria might be considered a single 'species' and show that individual genotypes are low in abundance, suggesting that individual genomes are insufficient fro inferring ecological function in the water collumn.

Contributions to Other Disciplines:

We believe our findings will specifically impact the fields of microbiology, ecology, evolution and genomics. Microbiology has lagged behind in formulation and testing of evolutionary and ecological theories. For example, due to a lack of well-constrained diversity estimates it has not been possible to put observations on microbial community structure in the context of the rich body of literature assembled for animals and plants. We believe that our contributions will provide stimulus to compare community patterns among all forms of life. Furthermore, our data provide an important framework and context for interpretation of current effort to access microbial diversity in complex communities via environmental genomics. The work has produced an extensive database that is currently being used by other investigators to test theories of species diversification and community assembly. Software products developed within this project have application in many areas of bioinformatics.

Contributions to Human Resource Development:

This project has contributed to the training of students at several educational levels including both underepresented minorities and women, and has been the focus of outreach efforts to secondary and high school teachers and local community leaders. Thus this effort has helped to prepare students for career oportunities in research and has increased awareness among lay public and educators of the importance of marine microbiology to the broader society.

Contributions to Resources for Research and Education:

Approximately 800 environmental 16S rRNA and HSP60 gene sequences have been submitted to Genbank. The addition of this unique data set to this public datatbase will be broadly valuable for research and education.In addition numerous gene libraries and more than 350 bacterial isolates have been collected and will be made available to the public through deposit to the Ocean Genome Legacy Marine Genome Resource.

Contributions Beyond Science and Engineering:

The microbial populations of the oceans strongly influence human health and welfare through thier contribution to food production, climate modification, distribution of pathogens and a variety environmental services(waste treatment, oxygen production, mineral cycling, etc.) The affects of human activities on such microbial processes have largely been a matter of conjecture since an insufficient theoretical framework has been available to apply and test concepts of modern theoretical ecology. The work done here makes inroads toward a more complete

understanding of microbial community diversity, assembly and dymanmics; knowledge that will be necessary to inform future environmental regulatory policy.

Categories for which nothing is reported:

Organizational Partners

Any Book