#2335 Topological Data Analysis Of PhyloChip Assay Hybridization Scores Reveals Community Shift In Deep-Sea Oil Plume

Authors

T. Z. DeSantis[1,2], P.Y. Lum[3], Y. Piceno[1], E. Dubinski[1], L. Tom[1], P. Hu[1], G. Singh[3], G. Carlsson[3], G. Andersen[1]

[1]Lawrence Berkeley National Laboratory, Berkeley, CA
[2]Second Genome, Inc., San Francisco, CA
[3]Ayasdi, Inc., Palo Alto, CA.

ABSTRACT

Background: The 2010 oil well blowout in the Gulf of Mexico was the deepest and one of the largest oil spills in history. We hypothesized that distinct communities would exist within the resulting plume compared to pristine waters at similar depths and that changes in specific populations can be rapidly detected with the PhyloChip Assay (Second Genome, Inc.) followed by Topological Data Analysis (Ayasdi, Inc).

Methods: Amplicons of 16S rRNA were generated from 17 water samples collected in and near the plume. The overnight PhyloChip Assay was performed to collect hybridization fluorescence data from over one million probes. TDA was performed with a novel software package, Iris (Ayasdi Inc.) using either the most varying probes or all probes to find topologically distinct domains useful for sample classification. In both procedures, probe responses distinguishing the sample types were evaluated for taxonomic inference.

Results: The water samples were classified into 3 topologically distinct groups: plume, non-plume, and boundary. Using both methods probes, low diversity was found in the plume compared to the non-plume. Probes that best discriminate the in-the-plume group from the out-of-the-plume group were identified. Within the plume, we observed a general decrease in Verrucomicrobia, Gammaproteobacteria, Flavobacteriales, and Prochlorococcus while Oceanospirillales and Pseudoalteromonadales displayed elevated populations. Furthermore, multiple probes without a match in the current Greengenes 16S rRNA database displayed significantly different hybridization among the sample types.

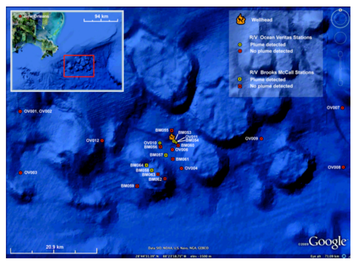
Conclusions: PhyloChip-TDA rapidly revealed the community shift within the oil plume. The focus on individual probe responses as opposed to traditional "probe sets" allowed discovery of topological substructures, confirmed population shifts from previous methods and tracked potentially novel strains of bacteria that were enriched in the oil plume.

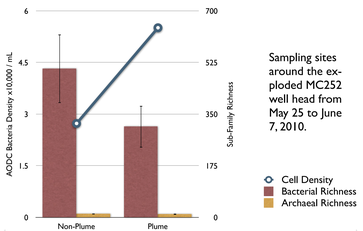
INITIAL RAPID ANALYSIS

Deep-Sea Oil Plume Enriches Indigenous Oil-Degrading Bacteria

Terry C. Hazen,[1]* Eric A. Dubinsky,[1] Todd Z. DeSantis,[1] Gary L. Andersen,[1] Yvette M. Piceno,[1] Navjeet Singh,[1] Janet K. Jansson,[1] Alexander Probst,[1] Sharon E. Borglin,[1] Julian L. Fortney,[1] William T. Stringfellow,[1,2] Markus Bill,[1] Mark E. Conrad,[1] Lauren M. Tom,[1] Krystle L. Chavarria,[1] Thana R. Alusi,[1] Regina Lamendella,[1] Dominique C. Joyner,[1] Chelsea Spier,[2] Jacob Baelum,[1] Manfred Auer,[1] Marcin L. Zemla,[1] Romy Chakraborty,[1] Eric L. Sonnenthal,[1] Patrik D'haeseleer,[1] Hoi-Ying N. Holman,[1] Shariff Osman,[2] Zhenmei Lu,[2] Joy D. Van Nostrand,[2] Ye Deng,[2] Jizhong Zhou,[1,2] Olivia U. Mason[1]

8 OCTOBER 2010 VOL 330 **SCIENCE** www.sciencemag.org



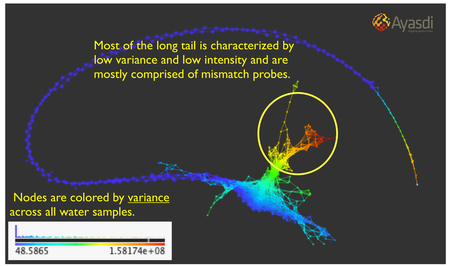


Sampling sites around the exploded MC252 well head from May 25 to June 7, 2010.

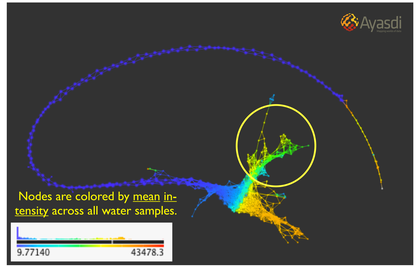
Hydrocarbon Increase Above Background:
•EPA Gulf of Mexico Hydrocarbon Concentration Allowance* = 29,000 parts per billion (29,000 mg/L)
•Deep Horizon Oil Spill Plume Hydrocarbon Concentration = 139 parts per billion (00.139 mg/L)
*The EPA NPDES (National Pollutant Discharge Effluent Standard) permits for Offshore Gulf of Mexico installations contain a NOT TO EXCEED limit of 29 ppm (42 mg/L Oil parts per million) on a daily basis AND a NOT TO EXCEED limit of 29 mg/L per day (29 parts per million) on a monthly basis.

•Immediately after the spill, microbial community analysis of deepwater plume and non-plume samples was conducted by LBNL. Differences in quantitative composition of 16S rRNA gene sequences was measured by the PhyloChip Assay and contrasted using a weighted Unifrac distance matrix. Plume and non-plume communities were significantly different as determined by permutational analysis of variance ($p < 0.005$).

PROBE LEVEL TDA ANALYSIS OF PHYLOCHIP™ ASSAY



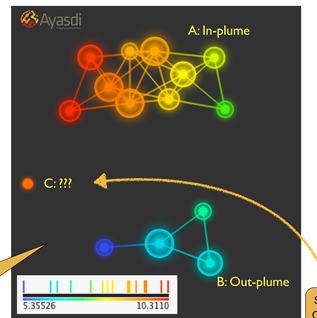
Metric: Euclidean; Geometric Lens: Gaussian Density Eq=true, Variance (Eq=false) Resolution: 150, Gain: 3.0X



- All 1,016,064 probes were analyzed across 17 samples in ~15 minutes using a single CPU core.
- Probes with similar response patterns across samples (Euclidean distance) are placed within a node, thus each node contains a subset of probes. Probes can belong to more than one node.
- Edge is drawn between nodes containing probes in common.
- The network output is a view of 656 nodes discovered. Probes with similar response patterns are next to each other in the image.
- 1,537 probes with both high means and high variance were subsequently passed to Method 1 (below) to categorize the oceanic samples are circled in yellow.

Are distinct microbiomes observed within the data set?

- Three topological search methods were used to study the set of samples:
  - **Method 1**: Restrict TDA sample clustering to only observations from the 1,537 probes with the highest means and highest variance.
  - **Method 2**: Allow TDA sample clustering to utilize all 1,016,064 probes.
  - **Method 3**: Same as 2, but substitute two of the PhyloChip results with replicates performed on different days with less gDNA template input into PCR.
- Topology overview
  - The field of mathematics concerned with the study of *shapes*.
  - The topological search method uses a distance metric and a set of *lenses* to construct a map of neighborhoods of the underlying data.
  - The shape of data is represented as a set of neighborhoods (denoted by nodes in the the networks) which are connected by edges.
  - The neighborhood construction procedure is:
    1. Compute the all-vs-all sample distance matrix using the cosine metric.
    2. Compute the density of each sample using a Gaussian-kernel density estimator. The density of a sample is proportional to the number of other similar samples in its vicinity (based on the Cosine metric).The density estimate is used as a lens for topological exploration.
    3. Next, the set of samples is ordered by density into 15 overlapping bins.
    4. Each bin is then clustered into a set of nodes and finally we draw edges between nodes which have samples in common.
    5. Run times for methods 1, 2 and 3 were <1min, ~5min and ~5min, respectively



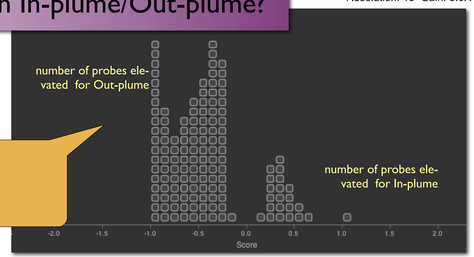
In all methods, **three** distinct clusters were observed.

Metric: Cosine Geometric Lens: Gaussian Density Resolution: 15 Gain: 3.0X Equalize: true

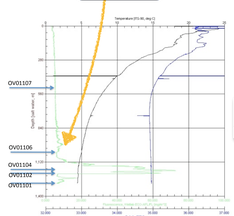
Sample OV01106 - sample from superior plume border, 1,042m

Which 16S rRNA probes distinguish In-plume/Out-plume?

- Omit boundary sample.
- Calculate a Kolmogoroff-Smiernov score for each probe by comparing the two distributions of the probes's intensities inside and outside the plume.
- Collect only most extreme K-S scores of 1 or -1 ($p = 0.005$).
- 140,330 probes collected.



Distribution Kolmogoroff-Smiernov scores.
Each histogram block represents 6784 probes.
Confirms loss of diversity in the plume compared to out of the plume.



Control Number: 4475 Session Title: Populations and Communities (Division N) Session Number: 204 Presentation: Tuesday May 24, 2011 10:45 AM - 12:30 PM Poster Board Number: 2335

1,016,064 probes in G3 PhyloChip Assay

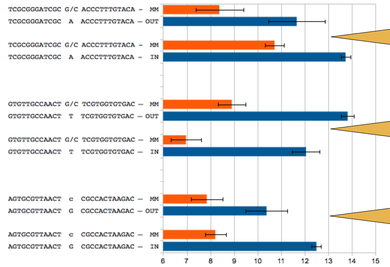
16S rRNA gene Probe Quartet
1. Top strand perfect match
2. Top strand mismatch
3. Bottom strand perfect match
4. Bottom strand mismatch

140,330 probes pass K-S test

27,842 *without* quartet controls
probes originally designed to be MM to database organisms could be tracking novel organisms.

122,488 *with* quartet controls

44,428 quartets PM>>MM(s)
mean response scores (r) found separately for grouped In-plume and Out-Plume samples. Filter set r >=0.95 in at least one group.



Hippea increase within plume

Tepidanaerobacter decrease within plume

Marinobacter increase within plume

- Annotation Approach
- 16S rRNA greengenes database with 2011 taxonomic annotations (McDonald, 2011, in review) was obtained.
- BLAST PM probe (-w 13 -b, -v 100K)
- BLAST MM probe(s) (-w 12 -b, -v 100K)
- PM must have alignment length >=24 and %id >=92 and under 100K HSPs
- PM score must be > MM score(s) for a subject to "hit"
- Annotating hits
  - 5 points awarded to subject id for perfect PM match
  - 1 point awarded to subject id for imperfect PM match
  - Determine the % of points awarded to each species ranking.
  - If >=90% is achieved, then node is associated with quartet.
  - Else, move one node toward trunk and repeat.

CONCLUSIONS

- The PhyloChip assay quantitatively tracks changes in microbial populations
- TDA is novel, sensitive and fast mathematical approach to analyzing complex, large metagenome data sets.
- The plume and non-plume bacterial communities were significantly different.
- Reproducible, quantitative probe responses are obtained with hybridization allowing confident determination of molecular signatures

ACKNOWLEDGMENTS