Ocean Cloud Commons: A Cyberinfrastructure for Microbial Ecology

Bonnie Hurwitz[1] , Illyoung Choi[2], and John Hartman[2]

[1]Dept. of Agricultural & Biosystems Engineering [2]Dept. Of Computer Science; The University of Arizona, Tucson AZ, USA

Introduction

Overview: The Tara Oceans Expedition has provided the largest publicly available contiguous dataset available in genomics for any scientific project in the world. Using the research schooner Tara and modern sequencing and state-of-the-art imaging technologies, a multinational team of scientists sampled microscopic plankton at hundreds of sites and depths in all the major oceanic regions. The Tara Oceans Expedition data have been released, but it is a challenge for researchers to access, manipulate, and analyze such large-scale resources.

Building a Cloud Data Commons: This project creates an Ocean Cloud Commons (OCC), a cloud-based resource and repository allowing researchers to query the Tara Oceans Expedition Data in the cloud; it also makes available comparative metagenomic tools through the Ocean Treasure Box (OTB).

Cyberinfrastucture Partnerships: The Ocean Cloud Commons and Ocean Treasure Box build upon established partnerships with organizations such as CyVerse Cyberinfrastructure, Agave Platform, OpenCloud, and computing facilities at the Texas Advanced Computing Center.

Global Comparisons: Taken together, the OTB tools and OCC data resources enable researchers to address global-scale questions about the distribution of microbes across the sea that affect climate and ecosystem function.

Scaling to Global Microbiology

Integrate large-scale –omics datasets

Interlink physiochemical and environmental context

Examine spatial scales across diverse ecosystems

Cooperate among disciplines to harmonize data

Maintain data on sampling and processing protocols

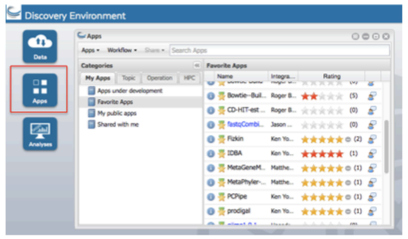
TARA OCEANS **Studying Global Oceans Systematically**

Integrating viruses, bacteria, archaea, protists, metazoans

Interweaving genomics, optics, physiochemical, satellite

A Cyberinfrastructure

CYVERSE **DISCOVERY ENVIRONMENT**



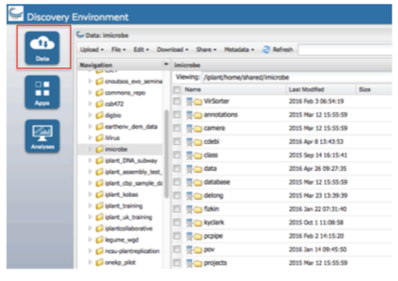
- ✓ A platform to run bioinformatics applications
- ✓ Integrate data and computation
- ✓ Build your own tools using Docker and HPC

Hundreds of tools available through a simple **web-based platform**

iMicrobe Metagenomics Apps

Discover Data	Implement Tools & Compute	Protocols & Community Network
iVirus http://ivirus.us		 protocols.io
iMicrobe http://imicrobe.us	CYVERSE	VERVENet

CYVERSE **DATA STORE**



- ✓ 100 GB initial allocation
- ✓ Automatic data backup
- ✓ Share data with collaborators
- ✓ Research data allocations available

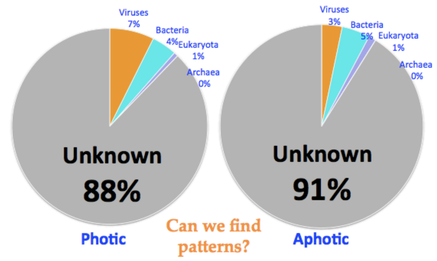
The resources you need to share and manage data with your lab, colleagues, and the community

iMicrobe Data Commons

Apps for Microbial and Virus Ecology

Vignette: Virus Ecology

The Vast Viral Unknown



New Big Data Algorithm

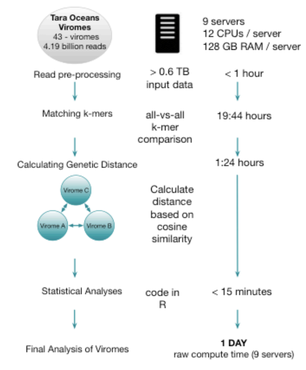
LIBRA Pipeline

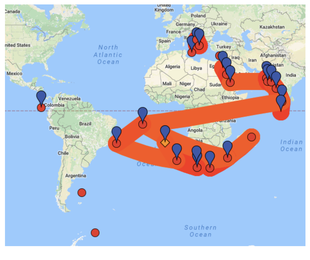
Hadoop Cluster: 10 servers

Compute time: 1 day

Data Used: 4.2 billion reads

Automated Pipeline: Enhanced reproducibility





Key Algorithm Components

Removes redundant k-mers and maintains read mapping

Partitions data equally between nodes for processing

Removes low abundance k-mers from sequencing error, contaminants or artifact

Implements a linear-time algorithm for an all-vs-all comparison of k-mers between all samples

Accounts for differences in k-mer abundance not just raw genetic similarity

Examining the Unknown

The 4C's – Connections, Counts, Context, Closeness

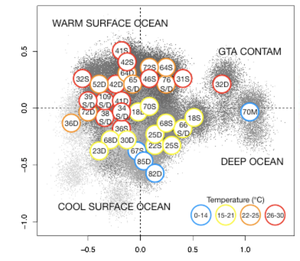
Unify data by connecting sequences

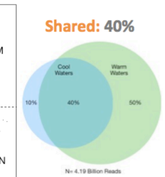
Count the frequency of connections

Learn the context in which samples are connected

Determine the closeness of the similarity based on context

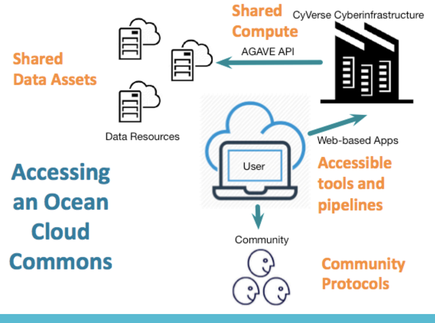
Temperature Defines Viral Community Distribution





Everything is Not Everywhere

Ocean Cloud Commons



Conclusions

Data "Clouds": are community-wide data assets from extremely large data sets

Big Data Analytics: to derive meaning from massive unstructured sequence datasets

Community Tools: can be developed to interrogate these data assets **iMicrobe**

Value: reveal patterns, trends, and associations in data that tell us about global biological patterns