JOHNS HOPKINS
UNIVERSITY

Zanvyl Krieger School of Arts and Sciences
Advanced Academic Programs

# 410.734.81 and 410.734.82
# Practical Introduction to Metagenomics

Topic: Gene clustering and data reduction

Instructor: Joshua Orvis

# Big Bad Data



It seems difficult to find a new genomics paper which doesn't have a requisite opening paragraph about how the scale and throughput of tools and techniques driving *-omics is increasing at a frightening pace. Indeed, the increase in almost all realms of biological data acquisition is exponential. (now I've done it too.)

Don't be afraid of your data. There are always advances in hardware, algorithms and protocols to deal with ever-changing data scale increases, and this week's readings and presentation introduce a selection of them.

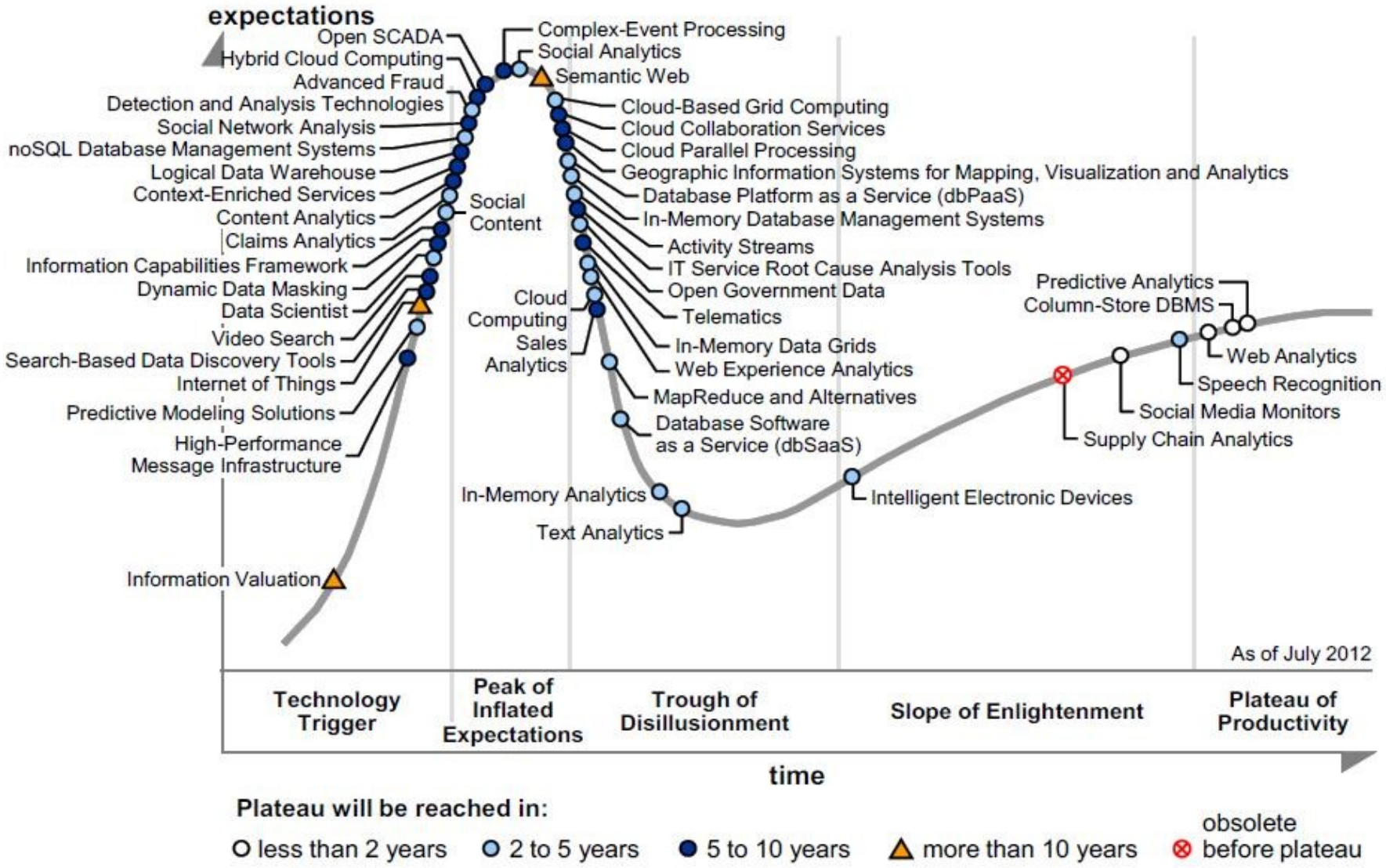
Topics include:

- File compression
- Read clustering
- Protein clustering
- Digital normalization

# Gartner Hype Cycle (just for fun)



Figure 1. Hype Cycle for Big Data, 2012




Source: Gartner (July 2012)

# Paper: Clustering algorithms



## Ultrafast clustering algorithms for metagenomic sequence analysis.

Li W, Fu L, Niu B, Wu S, Wooley J.
Brief Bioinform. 2012 Jul 6.
PMID: 22772836

The first of your two assigned publications this week is a survey covering the extensive uses of clustering techniques in metagenomics.

"By sequence clustering, a large redundant data set can be represented with a small non-redundant (NR) set, which requires less computation. Errors can be identified, filtered or corrected by using consensus from sequences within clusters. In addition, many fundamental questions in metagenomics can be readily addressed by clustering, such as the identification of gene families and the classification of species in a population."

Briefings in Bioinformatics Advance Access published July 6, 2012
BRIEFINGS IN BIOINFORMATICS. page 1 of 13 doi:10.1093/bib/bbs035

# Ultrafast clustering algorithms for metagenomic sequence analysis

*Weizhong Li, Limin Fu, Beifang Niu, Sitao Wu and John Wooley*

Submitted: 7th March 2012; Received (in revised form): 30th May 2012

**Abstract**

The rapid advances of high-throughput sequencing technologies dramatically prompted metagenomic studies of microbial communities that exist at various environments. Fundamental questions in metagenomics include the identities, composition and dynamics of microbial populations and their functions and interactions. However, the massive quantity and the comprehensive complexity of these sequence data pose tremendous challenges in data analysis. These challenges include but are not limited to ever-increasing computational demand, biased sequence sampling, sequence errors, sequence artifacts and novel sequences. Sequence clustering methods can directly answer many of the fundamental questions by grouping similar sequences into families. In addition, clustering analysis also addresses the challenges in metagenomics. Thus, a large redundant data set can be represented with a small non-redundant set, where each cluster can be represented by a single entry or a consensus. Artifacts can be rapidly detected through clustering. Errors can be identified, filtered or corrected by using consensus from sequences within clusters.

**Keywords:** *clustering; metagenomics; next-generation sequencing; protein families; artificial duplicates; OTU*

## INTRODUCTION

Metagenomics [1, 2] is a genomic approach that uses culture-independent sequencing to study the microorganism populations under different environments. It offers unprecedented vision of the identities, composition, dynamics, functions and interactions of the diverse microbial world and has become an important tool in many fields such as ecology, energy, agriculture and medicine.

Earlier metagenomics projects, such as Sargasso Sea [3], human gut [4] and soil [5], relied on traditional Sanger sequencing technology, so most of these projects have limited throughput. In recent years, the rapid advances of next-generation sequencing (NGS) technologies [6], such as 454, Illumina, SOLiD, PacBio and Ion Torrent, dramatically propelled the expansion of metagenomics research, and large 'waves' of metagenomics sequencing projects were launched to study a diverse range of microbial communities in their environments, such as the virome [7], farm animals [8] and the human microbiome [9, 10]. It is widely expected that many more environmental and microbiome samples will be studied by NGS technologies. However, the intrinsic

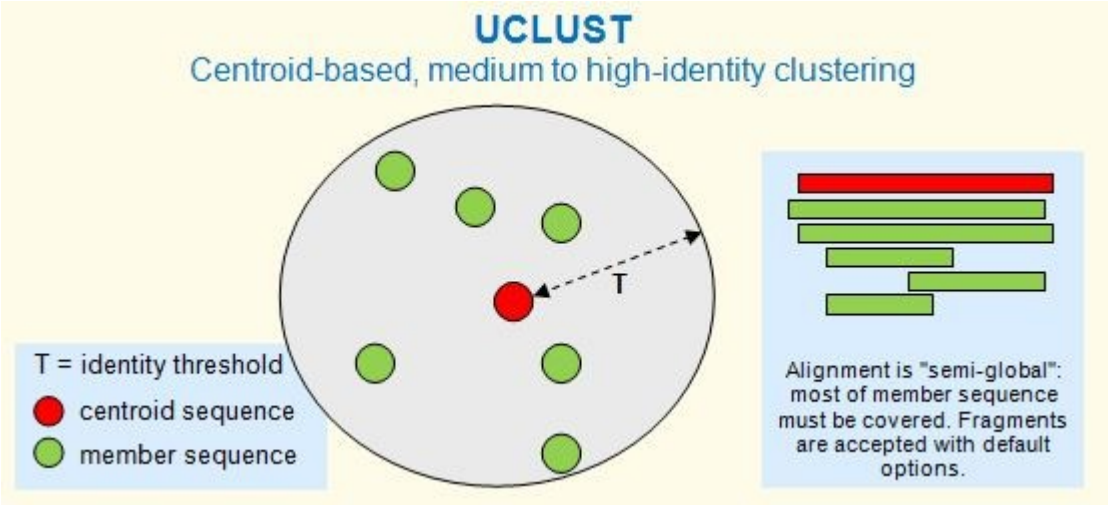
Corresponding author. Weizhong Li. Center for Research in Biological Systems, University of California San Diego, La Jolla, CA 92093, USA. Tel: 858-534-4143; Fax: 858-246-0644; E-mail: liwz@sdsc.edu
**Weizhong Li** is an Associate Research Scientist at the Center for Research in Biological Systems at University of California San Diego. Dr. Li has a background in computational biology. His research focuses on developing computational methods for sequence, genomic and metagenomic data analysis.
**Limin Fu** is a Postdoctoral Associate at the Center for Research in Biological Systems at University of California San Diego. Dr. Fu's background is mathematics. His research focuses on bioinformatics algorithm development.
**Beifang Niu** is a Postdoctoral Associate at the Center for Research in Biological Systems at University of California San Diego. Dr. Niu was trained as a computer scientist. His research focuses on next-generation sequence analysis.
**Sitao Wu** is a Staff Scientist at the Center for Research in Biological Systems at University of California San Diego. Dr. Wu has a background in electric engineering. His research interests include protein structure prediction and metagenomics.
**John Wooley** is a Professor of Pharmacology and Associate Vice Chancellor, Research at the University of California San Diego, as well as a member of the Center for Research in Biological Systems and the California Institute of Telecommunications and Information Technology. Dr. Wooley's background is biophysics and his current research interests include structural genomics and metagenomics.

# Clustering: UCLUST



Like so many algorithms you are exposed to in this field, it's important to not treat them as a magical black box that just generates results. Go beyond reading the options list and understand at least the basics of how it works. Also, know its limitations.

UCLUST, for example, is a very fast and useful tool for protein or nucleotide sequence clustering.




In short, it works by looking at the first entry in your sequence file and creating a 'centroid' sequence – the representative sequence of each cluster. Then, each successive sequence is then compared with all centroids. If it matches above your defined cutoffs it joins the cluster, else it becomes the centroid for a new one.

A benefit of this approach is that all sequences are searched only against the centroids or 'seeds' of clusters rather than all cluster members. This greatly increases the speed, but also creates an input order bias. The results you get can very much depend on the order of the sequences within your input file.

To at least control and document this, it's usually advised to sort the sequences in your file in descending size order before clustering.

# Digital normalization (diginorm)

JOHNS HOPKINS UNIVERSITY

410.734.81
Practical Introduction to Metagenomics

## A reference-free algorithm for computational normalization of shotgun sequencing data

C. Titus Brown, Adina Howe, Qingpeng Zhang, Alexis B. Pyrkosz, Timothy H. Brom
arXiv:1203.4802. [v2] Mon, 21 May 2012

This paper is having a significant impact on many different fields in genomics. It provides a possible solution to the "Big Data" problem by not only reducing many sequence datasets but also improving it at the same time.

Loaded with talk of billions of k-mers and millions of reads, it's also not the easiest thing to absorb immediately. The improvement, correction and reduction is certainly an emergent property of the algorithm, so you should try to use all the material available until you feel you've understood it. (see next slide)

### A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data

C. Titus Brown[1,2,*], Adina Howe[2], Qingpeng Zhang[1], Alexis B. Pyrkosz[3], Timothy H. Brom[1]
**1 Computer Science and Engineering, Michigan State University, East Lansing, MI, USA**
**2 Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI, USA**
**3 USDA Avian Disease and Oncology Laboratory, East Lansing, MI, USA**
* **E-mail: ctb@msu.edu**

#### Abstract

Deep shotgun sequencing and analysis of genomes, transcriptomes, amplified single-cell genomes, and metagenomes has enabled investigation of a wide range of organisms and ecosystems. However, sampling variation in short-read data sets and high sequencing error rates of modern sequencers present many new computational challenges in data interpretation. These challenges have led to the development of new classes of mapping tools and *de novo* assemblers. These algorithms are challenged by the continued improvement in sequencing throughput. We here describe digital normalization, a single-pass computational algorithm that systematizes coverage in shotgun sequencing data sets, thereby decreasing sampling variation, discarding redundant data, and removing the majority of errors. Digital normalization substantially reduces the size of shotgun data sets and decreases the memory and time requirements for *de novo* sequence assembly, all without significantly impacting content of the generated contigs. We apply digital normalization to the assembly of microbial genomic data, amplified single-cell genomic data, and transcriptomic data. Our implementation is freely available for use and modification.

#### Author Summary

#### Introduction

The ongoing improvements in DNA sequencing technologies have led to a new problem: how do we analyze the resulting large sequence data sets quickly and efficiently? These data sets contain millions to billions of short reads with high error rates and substantial sampling biases [1]. The vast quantities of deep sequencing data produced by these new sequencing technologies are driving computational biology to extend and adapt previous approaches to sequence analysis. In particular, the widespread use of deep shotgun sequencing on previously unsequenced genomes, transcriptomes, and metagenomes, has resulted in the development of several new approaches to *de novo* sequence assembly [2].

There are two basic challenges in analyzing short-read sequences from shotgun sequencing. First, deep sequencing is needed for complete sampling. This is because shotgun sequencing samples randomly from a population of molecules; this sampling is biased by sample content and sample preparation, requiring even deeper sequencing. A human genome may require 100x coverage or more for near-complete sampling, leading to shotgun data sets 300 GB or larger in size [3]. Since the lowest abundance molecule determines the depth of coverage required for complete sampling, transcriptomes and metagenomes containing rare population elements can also require similarly deep sequencing.

The second challenge to analyzing short-read shotgun sequencing is the high error rate. For example, the Illumina GAII sequencer has a 1-2% error rate, yielding an average of one base error in every 100 bp of data [1]. The total number of errors grows linearly with the amount of data generated, so these errors usually dominate novelty in large data sets [4]. Tracking this novelty and resolving errors is computationally expensive.

# Diginorm resources



The author, Dr. C. Titus Brown, pushes this technique with almost religious fervor, so we benefit from many different descriptions of it for different audiences. You should go through these to understand it as much as possible. In these he also expands the discussion to deal with metagenomic sequence specifically.

What is digital normalization, anyway?
http://ivory.idyll.org/blog/what-is-diginorm.html

Diginorm paper submission cover letter to PLoS One
http://ged.msu.edu/downloads/2012-diginorm-plos-cover-letter.pdf

Slides from Oct 2011:
"Scaling metagenome assembly"
http://www.slideshare.net/c.titus.brown/scaling-metagenome-assembly

Official diginorm paper site
http://ged.msu.edu/papers/2012-diginorm/

"Scaling metagenome sequence assembly with probabilistic de Bruijn graphs"
Explains Bloom filters
http://arxiv.org/abs/1112.4193v3

## Sequence compression



In the Unix/Linux world the *de facto* standard methods for compressing files is to use the Gzip or Bzip2 utilities. These are great general-purpose tools and provide us with a guide of the list of basic features any compression tool should have:

- Globally available on almost all modern Unix/Linux systems
- Very simple to use (basic execution with no or few required options)
- Accepts piped input (streams vs. files)
- Can be read by most programming languages natively without pre-decompression.

Notice that none of these actually have anything to do with the content of the data involved.

Something built for our data specifically should be able to run faster and with a much higher compression ratio.

Now the developers just needed to be motivated.

This year there was a contest with a $15,000 prize for the “best novel open-source NGS compression algorithm submitted before the closing date of 15 March 2012.”

The winner was James Bonfield of the Sanger Institute, but the full list of entrants along with the statistics of their tools on a controlled dataset is found here:

http://www.sequencesqueeze.org/

This is a great resource, but the details about a lot of these are sparse and not peer reviewed. If you're more interested in published tools, there are several available.

When you evaluate these, take care to remember the list of 'requirements' on the previous page to see if it meets them all. The last one, particularly, is important for adoption and often the most overlooked. I give tools a lot more weight if the author has taken the time to include modules for languages like perl, python, etc. so that they can be streamed without pre-decompression.

## Compression comparison



Here is a quick test of compression with gzip and bzip2 using a larger HMP sample set.

### Input description

- format: FASTQ
- reads: 213,716,553
- bases: 21,799,088,406
- avg length: 102.0

### Results

| method | compression time | decompression time | compressed disk usage | fold compression |
|---|---|---|---|---|
| gzip | 138m12.459s | 13m18.816s | 21726992613 | 2.58 |
| bzip2 | 123m12.113s | 37m39.905s | 15000474672 | 3.74 |
| beetl | n/a | n/a | n/a | n/a |
| dsrc | 40m27.242s | 22m10.261s | 13846262759 | 4.06 |
| quip | 48m9.454s | 53m19.088s | 11383703657 | 4.93 |

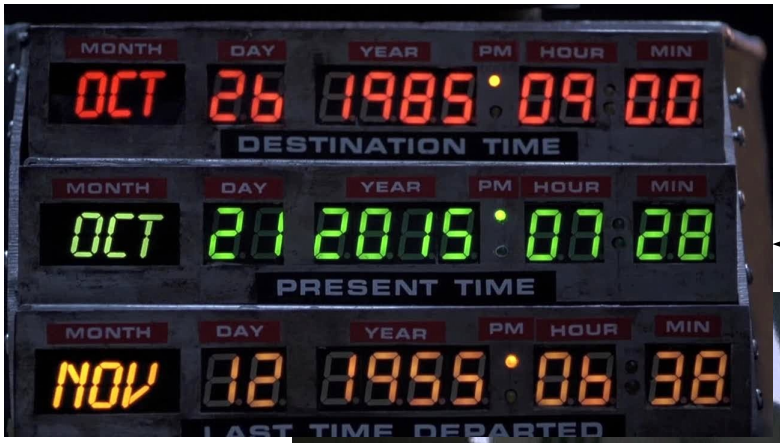
DSRC has the fastest compression time, significantly faster than gzip and bzip2 while at the same time reducing the overall FASTQ file size by 37% compared with gzip, which is what the Illumina machines use by default.

What's more, DSRC is distributed with python libraries to read the compressed files directly.

There are many great alternatives to gzip/bzip2 for lossless NGS sequence compression.

## Scientists: We only have 3 years left






## Further reading

The gzip INFLATE and DEFLATE algorithms
http://www.gzip.org/algorithm.txt

Compression of DNA sequence reads in FASTQ format
http://bioinformatics.oxfordjournals.org/content/27/6/860.full

Search and clustering orders of magnitude faster than BLAST (UCLUST & USEARCH)
http://bioinformatics.oxfordjournals.org/content/26/19/2460