Making the Dark Matter of Biology Visible

Joel Berendzen, P-21;
Mira Bussod, T-6;
Judith Cohn,
Nick Hengartner, CCS-3;
Ben McMahon, T-6

The advent of large-scale shotgun sequencing of entire microbial communities, viewed as a meta-organism, promises to make the dark matter of biology visible by helping us understand microbial communities important to medicine, ecology, industry, or agriculture [1]. The basic problem in using shotgun sequencing data from environmental samples is to roughly identify the phylogenetic origin and possibly the gene each sequence read, often only 100–200 bp long, belongs too. The current approach is to compare each read to the entire library of sequenced data. While Genbank contains a truly enormous quantity of genetic sequence data to which comparisons can be made (over 1,500 distinct bacterial species, along with approximately 100 fungi, 100 archaea, 50 metazoa, and 50 plants and protozoa), its coverage of organisms on the tree of life is uneven. A cursory identification of genomic fragments can be made by assessing its similarity to elements in the database, using versions of the blast algorithms developed at LANL in the 1980s. This comparison is computationally cumbersome because it may take several weeks on large computer clusters. This computational burden is the main bottleneck for extracting information from the shotgun sequencing of environmental samples.

A game-changing approach that enables us to analyze the genomic content from environmental samples is to look for meaning instead of similarities in the sequence reads. The elements required for success in analyzing languages are also present in the genetic language of DNA. A typical bacterial genome encodes for several thousand proteins, with a total of around one million amino acids involved in their sequences. Like the English language, which constructs its entire literary cannon with only 26 letters, nature encodes the genomes of all of its great creatures with the same 20 amino acids. These 20 amino acids can conveniently be represented with one-letter codes: A for alanine, C for cystine, D for aspartic acid, etc. If a genome is analogous to a book, and an amino acid is analogous to a letter, a paragraph would likely correspond to a protein, and a chapter to a group of related proteins, known as an operon.

To define the concept of a genomic word, a basic element that has meaning for many organisms consider Fig. 1, which shows how the same sentence from different languages share n-grams that can be used to associate meaning with words. This observation has been exploited by Google translate. For biology, shared k-mers of amino acid play the same role. Figure 2 shows a histogram of the number of amino acid exact matches between *E. coli* and a variety of other bacteria as a function of the length of this exact match. Short patterns of amino acids are ubiquitous, with 300 million instances of matching amino acid 3-mers between the divergent organisms *E. coli* and *B. subtilis*. They are too numerous to help convey meaning. On the other hand, exact matches of 20-mers of amino acids occur mainly in groups of closely related organisms. They are useful in identifying organisms in that group, but cannot be used to learn about common functionalities.

English	**Science is the great antidote to the poison of superstition.**
French	**La Science est le grand antidote au poison de la superstition.**
Spanish	**La ciencia es el gran antidoto al veneno de la supersticion.**
Dutch	**Wetenschap is het grote tegengif aan het vergift van bijgeloof.**
German	**Wissenschaft ist das grosse Antidot zum Gift von Aberglauben.**

Fig 1. The same sentence from different languages share n-grams that can be used to associate meaning to words.

There are typically 2,000 matching k-mers of length 10 between divergent organisms and over tens of thousands shared 10-mers for more closely related organisms. As a comparison, we expect one random match between any two genomes. Independently, David Baker identified 9-mers of amino acids as a relevant unit for protein structure prediction (rosetta). These shared 10-mers arise both through natural selection and inheritance and are thus expected to be present in organisms not in our database [2].

The recently published paper by the authors [2] exploits this idea and combines it with evolutionary theory and web search-engine technology to develop a software package Sequedex [3] that can classify raw metagenomics reads 250,000 times faster than the current pipelines.

This new tool makes possible a much more novel use of next-generation sequencing, more along the lines proposed by the National Academy study—that of identifying microbial communities through their phylogenetic profile. Figure 3 shows the rolled-up phylogeny of 242 environmental microbial communities obtained from 30 separate studies and obtained from National Center for Biotechnology Information Sequence Read Archive and the Community Cyberinsfrastructure for Advanced Microbial Ecology Research and Analysis website. Literature

For more information contact Nick Hengartner at nickh@lanl.gov.

Funding Acknowledgments
LANL, Laboratory Research and Development Program

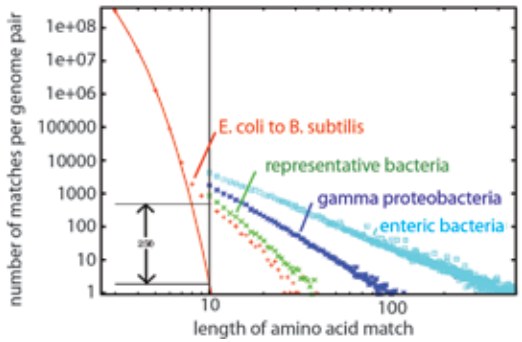

Fig. 2. Graph of the number of shared k-mers between E.coli and B. Subtilis as a function of k, on a log-log scale. Initially, the number of k-mers decrease exponentially. At k=8, the exponential decrease becomes polynomial (powerlaw). Extrapolation of the exponential decay to k=10 shows that the signal to noise ratio of 10-mers is 250.

references for each of the samples are available in the online documentation for the Sequedex software package [3]. The community similarities computed from signature-peptide-based distance metrics, such as that shown in Fig. 2, are much more stable and predictive than those determined by ribosomal RNA (rRNA) surveys because they are composed of the occurrence of thousands of conserved elements, rather than the variable portion of the ribosome.

The Sequedex algorithm can be adapted to design viral and bacterial pan-diagnostics using shotgun sequencing. Figure 4 shows a histogram of viral reads identified from 100 million reads sequenced from an RNA sample preparation from a clinical diarrheal sample. The y-axis shows the number of distinct reads identified, plotted against the node-number of the RNA virus portion of a one-per-species tree of viruses. Four distinct peaks show up against an essentially flat baseline. Insets show the viral phylogeny in the region where matching reads were identified, including adeno-associated virus, resperatory syncytial virus (RSV), sapovirus, and human enterovirus. The analysis, using Sequedex, took less than 15 minutes on a laptop. Competing methods [such as Basic Local Alignment Search Tool (BLAST)] applied to amino-acid translations and compared to the non redundant (NR) protein database require approximately one week of computing time on a cluster of computers, making the process unsuitable for routine use. As already noted, the process does not rely on the virus being closely related to a virus in the reference database, only that some of the genes contain conserved regions. Even such distinct pathogens as measles, mumps, and Nipah virus share signature peptides in numerous places across their genomes, which opens the possibility for broad characterization of endemic pathogens in a given environment.

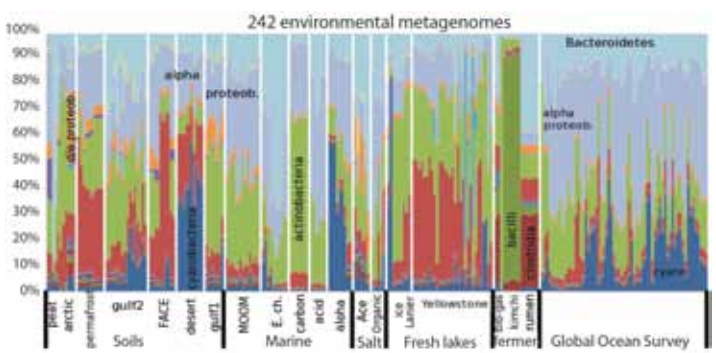

Fig. 3. Phylum-level phylogenetic profiles across 242 environmental metagenomic sample sets taken from 30 separate studies. Not only are distinctive profiles observable for the different sites, but field replicates from similar studies show a 99.9% similarity when comparing the full-resolution phylogenetic profiles to one another [2].

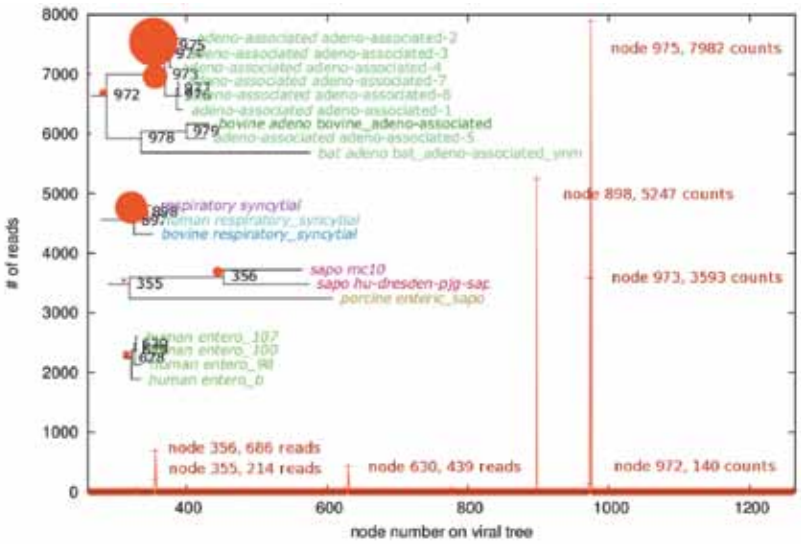

Fig. 4. Histogram of 100-bp reads identified in an RNA sequencing run of 100 million human, bacterial, and phage reads sequenced from 200-mL clinical diarrheal samples. Viral reads were identified in one hour of CPU time on a desktop computer, requiring 2 Gb of memory. Data were obtained from the laboratory of Charles Chiu, UCSF.

[1] Handelsman J., et al., "The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet," National Research Council, Washington, DC (2007).

[2] Berendzen, J.R. et al., *BMC Res Rep* **5**, 460 (2012).

[3] Sequex, http://sequedex.lanl.gov (2012).

Bibliography
Rohl, et al., *Meth Enzymol* **383**, 66 (2004).
Garten, et al., *Science* **325**, 197 (2009).
Korber, et al., *Science* **288**, 1789 (2000).
Ou, et al, *Science* **256**, 1165 (1992). http://hfv.lanl.gov