Topical: Data Science Applications for Planetary Protection

Lisa Guan[1,*], Ashish Mahabal[2], Chris Francis[3]

1 NASA Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA
2 California Institute of Technology, Pasadena, CA
3 Indian Institute of Technology, Gandhinagar, India

[*]Primary author contact: lisa.guan@jpl.nasa.gov, 818-354-2814

Background

Data science has become increasingly prevalent in the field of biology, a field full of complex and poorly understood systems. Recent advances in high throughput molecular biology techniques are capable of generating large amounts of data that can then be analyzed to reveal patterns within complex systems. This hypothesis-free and data-driven discovery approach of data science to build and test models, based on observed patterns rather than sometimes biased presuppositions, allows scientists to investigate and explore possible trends within their experiments that could never have been possible with traditional workflow of hypotheses followed by data-processing. Planetary Protection seeks to understand microbial life in extreme environments such as spacecraft assembly cleanrooms and spaceflight. To do so, it is headed towards applying molecular biology techniques to answering the questions that are needed for future missions.

Planetary protection requirements were first established during the Viking-era (1960-70s) and are dependent on culture-based assays that provide limited information. Only organisms that can be cultured in select laboratory conditions can be discovered, which leaves a vast majority of microorganisms undetected, particularly those that are most likely to adapt to harsh conditions, survive cleaning procedures, and are unlikely to grow in mild lab conditions.

Future missions with PP requirements that assess risk of certain types of organisms will require more than just culture-based spore data. For these missions, extremophile organisms are the most concerning and are of greatest interest.

A leading bioassay candidate for identifying organisms that can't be cultured is metagenomics, which uses DNA or RNA sequencing of the entire environmental sample to get a broad profile of all organisms that are present. To do this, nucleic acids are extracted from a spacecraft or spacecraft-associated surface, deep-sequenced with high-throughput methods, and the resulting data is analyzed for taxonomic or functional identification.

Genomic data, or any type of -omic data, which is focused on generating biological data at a deep level and for the entire community of microorganisms, can quickly become voluminous (gigabytes per sample) and requires scalable computational and statistical methods to process.

While software (including a large number of open-source programs) exist to process genomic data, finding new, fundamental insights within our specialized dataset would best be done by statistical models and machine learning algorithms tailored to the challenges that NASA faces. We need to build models of biological observations, repeatedly test, and validate them to recognize patterns and extract meaningful inferences. Nuances of next-generation molecular biology methods don't yield black-and-white results that can clearly answer Planetary Protection requirements.

Current application of data science for PP

Challenges with metagenomics

While using metagenomics to identify spacecraft microbes has been something that PP has been exploring for the past decade [1–3], there remain challenges that need to be addressed prior to implementing it as a bioburden assay. Specifically, DNA sequencing of low-biomass samples is prone to contaminant DNA from sources such as laboratory environments and reagents. The resulting raw data from low-biomass samples is unavoidably noisy, with the contaminant DNA often overwhelming and confounding results. When analyzed with metagenomic classifiers, even extraction controls and no-template controls can contain several taxa of bacteria [4, 5].

The types of surfaces that are sampled for Planetary Protection purposes are typically spacecraft hardware surfaces that have undergone cleaning procedures and can be considered "ultra-low" biomass in the range picograms to nanograms of total DNA in each sample. The presence of contaminant DNA and sequencing noise in these samples would skew results and could falsely identify bacterial contamination on hardware and risk the hardware not meeting Planetary Protection requirements for flight. Reliable and reproducible quantitative methods need to be implemented on low-biomass metagenomic data to minimize noise and remove false positives in identified taxa. Only then can metagenomics, as a highly informative and sensitive technique, be used for flight project cleanliness guidelines.

How Data Science is used to address that challenge

In an effort to address the false positives that are prevalent in low-biomass metagenome samples, there has been a recent effort at JPL to apply data science techniques to clean up contaminants post-hoc. A pilot project was funded by the JPL Data Science group aimed at distinguishing true positive taxa in metagenomic samples from false positives. The current generation of DNA sequencers produce millions of short sequences, called reads. The project attempts to utilize genome coverage, the number of sequencing reads mapped to a position in a known genome, as a way to estimate the likelihood that a read was generated by an intact microbial cell. While the average genome coverage is frequently used within algorithms in computational genomics, the complete information available in coverage profiles is currently not exploited to its full extent. For Planetary Protection compliance, only intact cells are considered; DNA fragments are not. Genome Coverage Profiles (GCP) of intact organisms follow a Poisson distribution [6], while GCPs of any laboratory contamination, residual fragments of DNA or organisms, do not. Therefore, the GCP of an organism should be significantly different when it is truly present in a sample (as an intact cell) vs. when a fragment of its DNA is present as contamination. The pilot project's main goal was to leverage this contrast in the GCP pattern to develop a machine learning model that can effectively distinguish between signal and noise, i.e. DNA from intact cells vs. DNA fragments from reagents. Without machine learning, there would be limitations to applying such a GCP-based approach as actual environmental samples typically contain upwards of 1000 different organisms, making manual comparisons for each genome impractical. Therefore, an automated, unbiased approach is needed to apply a GCP-based classification strategy for real-life application.

The methodology for the pilot project started with analyzing positive controls of a model microbial community in 1) several DNA concentrations to understand the effect concentration has on the ability to distinguish signal from noise and 2) a single concentration that were repeatedly sequenced with each batch of spacecraft DNA samples as a standard across the numerous batches. It was necessary to start with positive controls containing 10 known species in a mixed microbial community in order to generate GCPs for both known true organisms and known contaminants/noise. These were used to train and test various machine learning models.

GCPs were made by first gathering the genomes of the top 20 most abundant species within each sample (or in the case of positive controls, the 10 expected species and 10 most abundant contaminants). The most abundant species were determined using a Kraken2 and Bracken pipeline with default parameters [7, 8] and includes all identified taxa, including potential false positives. Well-assembled high quality genomes collected from RefSeq [9] of the 20 most prominent species were concatenated into one file and used as a reference genome against which the raw sequencing reads of the sample were mapped using BWA-mem [10]. This is done to find and filter out reads that map to more than one genome with equal quality (multi-mappers). Reads that randomly map to several genomes were removed as they are potential sources of noise because they may be mapping to highly conserved regions shared by many taxa. The remaining reads were mapped again to individual reference genomes from the top 20 list using BWA-mem [10] to generate coverage information at each position of the genome.

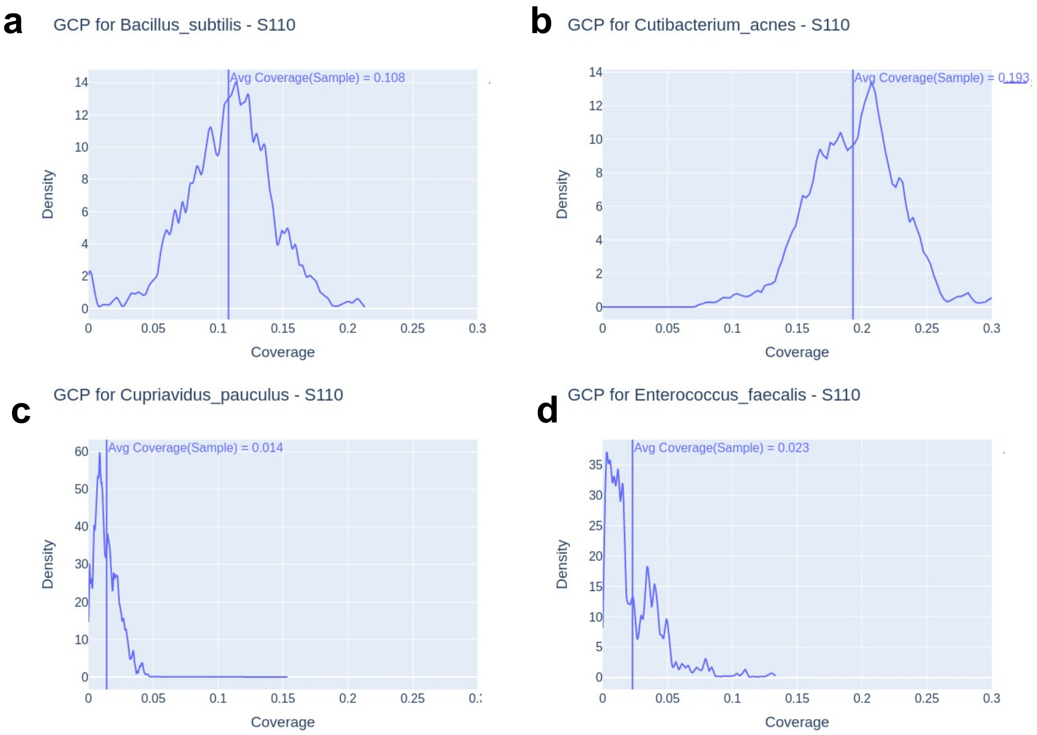

Figure 1: **Genome coverage profiles from a positive control sample of known input. (a)** and **(b)** GCPs from true positives. **(c)** and **(d)** GCPs from noise

The coverage profile was passed as input to the models in the form of a sequence of 100 numbers. Genome coverage plots were generated using Plotly [11]. Example GCPs are shown in Figure 1, both from true positives (Fig. 1a,b) and from noise (Fig. 1c,d). Various machine learning models that were tested included a Ridge Linear Classifier model, a One-Directional Convolutional Neural Network (CNN), and a Long Short-Term Memory (LSTM) model. Prelimary results using a small dataset for validation and testing show that intra-batch results are better than inter-batch results. Example with Ridge Linear Classifer model is shown in Figure 2. Continued testing and adjustments to the model are needed to confirm this method and improvements need to be made on several fronts. For example, the genome coverage profiles and the total coverage could be improved if fewer multi-mappers are removed. Removing multi-mappers is causing a steep drop in reads for many samples and is problematic since it reduces coverage information non-uniformly, resulting in lopsided values and non-quantitative results [12]. Longer reads or contigs from assembled reads could be used so that individual reads map to fewer organisms. Another approach could be to split the multi-mapping reads equally between each genome that it aligns to.

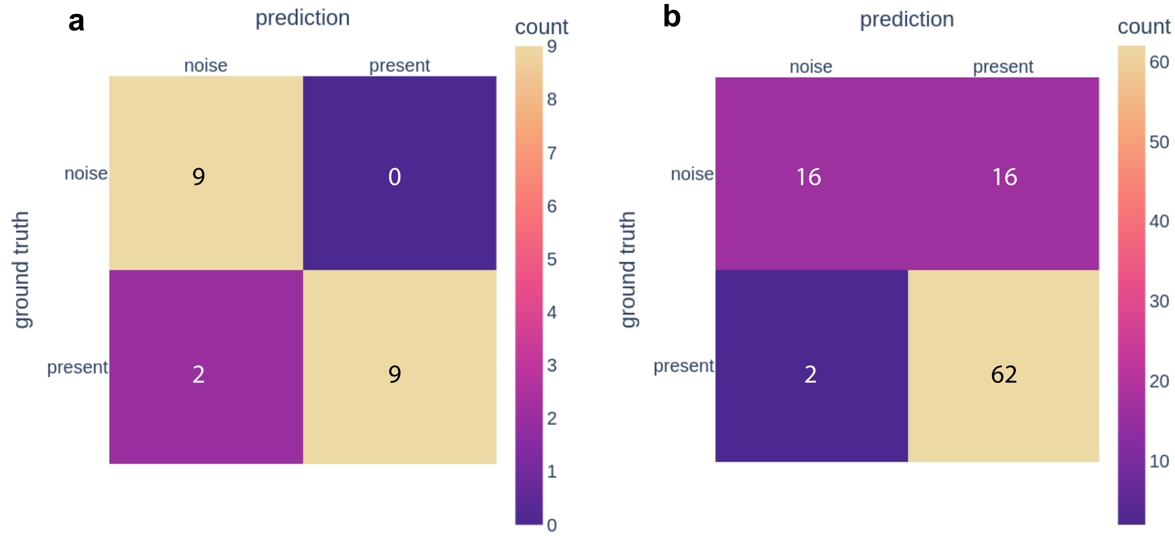

Figure 2: **Confusion matrices for Ridge Linear Classifier model preliminary results. (a)** Validation results: Accuracy = 0.9, Precision(present) = 1, Recall(present) = 0.818, F1 score = 0.9 **(b)** Test results: Accuracy = 0.8125, Precision(present) = 0.795, Recall(present) = 0.969, F1 score = 0.873

The main product of this pilot project has been a framework to take an aspect of metagenomic data (in this case the coverage distribution) and thoroughly explored it to search for informative patterns. These first steps pave the way for more detailed, complex and quantitative estimations of microbial populations, even at low biomass. Generating, analyzing, and testing the large amount of data required for this task with statistical rigor is only possible with data science and machine learning. Visual analysis of genome coverage plots by eye is subject to user bias and can be difficult to replicate. As biological research such as this relies more on big data, researchers will find themselves increasingly leaning on computation methods and data science. The hope is that integrating data science in the planetary protection discipline can equip the field with modern molecular biology techniques that will benefit future flight missions.

Acknowledgements

The research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration (80NM0018D0004).

References

1. Venkateswaran, K. *et al.* Molecular microbial diversity of a spacecraft assembly facility. *Systematic and Applied Microbiology* **24,** 311–320. doi:10.1078/0723-2020-00018 (2001).
2. Weinmaier, T. *et al.* A viability-linked metagenomic analysis of cleanroom environments: eukarya, prokaryotes, and viruses. *Microbiome* **3,** 62. doi:10.1186/s40168-015-0129-y (2015).
3. Danko, D. C. *et al.* A comprehensive metagenomics framework to characterize organisms relevant for planetary protection. *Microbiome* **9,** 82. doi:10.1186/s40168-021-01020-1 (2021).
4. Eisenhofer, R. *et al.* Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. *Trends in microbiology* **27,** 105–117. doi:10.1016/j.tim.2018.11.003 (2019).
5. Salter, S. J. *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology* **12,** 87. doi:10.1186/s12915-014-0087-z (2014).
6. Lindner, M. S., Kollock, M., Zickmann, F. & Renard, B. Y. Analyzing genome coverage profiles with applications to quality control in metagenomics. *Bioinformatics* **29,** 1260–1267. doi:10.1093/bioinformatics/btt147 (2013).
7. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biology* **20,** 1–13. doi:10.1186/s13059-019-1891-0 (2019).
8. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: Estimating species abundance in metagenomics data. *PeerJ Computer Science* **2017.** doi:10.7717/peerj-cs.104 (2017).
9. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research* **44,** D733–D745. doi:10.1093/nar/gkv1189 (2016).
10. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760. doi:10.1093/bioinformatics/btp324 (2009).
11. Plotly Technologies Inc. *Collaborative data science* 2015.
12. Deschamps-Francoeur, G., Simoneau, J. & Scott, M. S. *Handling multi-mapped reads in RNA-seq* 2020. doi:10.1016/j.csbj.2020.06.014.