Statistical Classification of Biosignature Information: Combining Elemental, Molecular, Reflectance, and Raman Data to Increase Life Detection Confidence

Diana Gentry[1], Abdullah Shahid[2], Tao Sheng[3], and Sunanda Sharma[4]
[1]NASA Ames Research Center (diana.gentry@nasa.gov), [2]North Carolina State University, [3]University of Pittsburgh, [4]Jet Propulsion Laboratory

Introduction: Planetary exploration missions seeking past or present signs of life carry not just a single instrument, but a suite. There is a need to study how these multiple data types can be combined to create "composite" biosignatures [1]. Algorithmic methods using existing data on living and non-living systems, though limited by the $n = 1$ of Earth, can nonetheless be informative. We assembled a database of 1277 measurements spanning 16 representative systems either *indicative* or *non-indicative* of life. Five classification (machine learning) methods were used on each individual data type, then on the entire set. This abstract summarizes the results; the data is described in more detail in [2], and methods in [3].

Methods: The four data types were elemental abundance (E), isotopic fractionation (I), VNIR reflectance spectra (V), and Raman spectra (R). Features within each type were chosen to reduce dependence on Earth-like biochemistry. Indicative samples included bacteria, leaves, humans, shells, and biogenic kerogens; non-indicative included meteorites, sand, and Mars regolith. Algorithms were k-nearest neighbors (KNN), logistic regression (LR), support vector machines (SVM), random forest (RF), Gaussian naïve Bayes (GNB), and a "voting" sum of the prior four. Principal component analysis (PCA) and $n - 1$ tests provided added introspection into the relative importance of each data type, sample, and feature.

Results: The combination of all four data types achieved ROC AUC $= 0.79$. Individually, elemental abundance had the highest AUC (0.8) and Raman least (0.71). Three data type subsets (E, E+V, E+I) outperformed the total set (E+I AUC $= 0.85$). When indicative samples were subclassed as *alive*, *non-alive*, or *mixed* (indicative + non-indicative), mixed samples were less accurately classified. PCA plots reflect the variety of effectiveness of the data combinations (Fig. 1).

In $n - 1$ tests, Mg abundance had the biggest positive impact on AUC ($+0.02$), followed by C, mean Raman peak width, Ti, and max Raman peak width ($+\text{AUC} > 0.01$); only mean Raman intensity and Cl had negative impact > 0.01 AUC. Carbonatite (non-indicative) was misclassified in a majority of cases; seawater and soil (mixed indicative) were misclassified in a significant minority.

Table 1: AUC for a subset of data type combinations.

	KNN	GNB	LR	RF	SVM	Vote
R	0.63	0.63	0.53	0.78	0.59	0.71
I+V+R	0.78	0.60	0.62	0.77	0.51	0.73
all	0.80	0.63	0.69	0.81	0.59	0.79
E+I	0.81	0.66	0.76	0.81	0.75	0.85

Conclusions: These results show that combining data types can result in novel "composite biosignatures" with significantly improved performance. However, more data is not always better; in several cases, specific data combinations decreased accuracy. Data types prone to similar misclassifications may be reinforcing errors.

Mixed samples are closest to plausible *in situ* samples, and their generally lower accuracy is an important caution. While feature selection for elemental and isotopic abundance was straightforward, features used for reflectance and Raman spectra were simply based on common tools; more advanced feature extraction methods could significantly improve these types' relative value. Beyond direct direction, secondary information like clustering and outliers can inform mission instrument selection or priority for *in situ* sampling and analysis.

This work assessed a limited number of samples and data types due to the sparsity of sources and labor needed to standardize it. These preliminary results are a strong proof of concept justifying further study.

References: [1] Neveu M. et al. (2018) *Astrobiology*. doi:10.1089/ast.2017.1773. [2] Sheng T. et al. (2024) (this conf.). [3] Shahid A. et al. (2024) (this conf.).

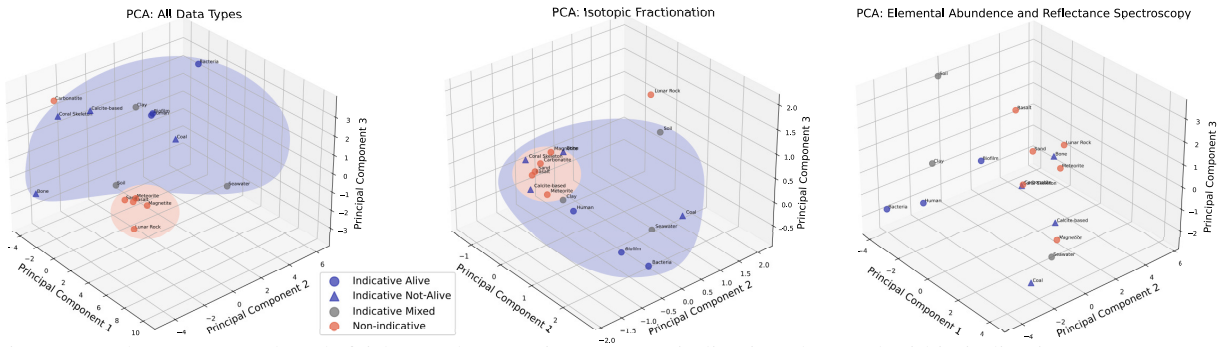

Figure 1: Select PCA results: (l) fairly good separation; (c) non-indicative clustered within indicative; (r) no pattern.