Fingerprinting Non-Terran Biosignatures:

Chemometrics and Complexity Patterns for Agnostic Life Detection on Ocean Worlds

Heather V. Graham[1], Sarah Stewart Johnson[2], Paul Mahaffy[1], Eric Anslyn[3], Andrew Ellington[3]

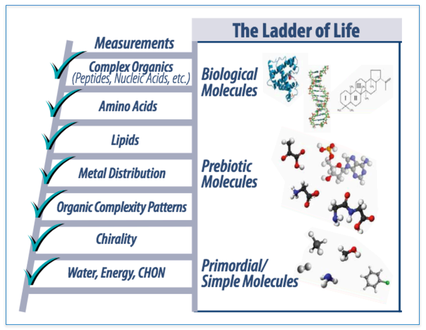
1) NASA Goddard Space Flight Center, 2) Georgetown University, 3) University of Texas

Agnostic Biosignature Detection

As we explore deeper into the Solar System, the likelihood of encountering organisms that do not share a common heritage with terrestrial life becomes much greater. Current life detection strategies rely mainly on identification of well-established and widely accepted biochemical features associated with terrestrial life.

How do we search for life as we *don't* know it?

Rather than looking for familiar presupposed molecular frameworks of Terran a better strategy would be to identify life by it's activity and general features An agnostic biosignature relies on attributes common to life instead of specific molecular building blocks[1].



The Ladder of Life describes biosignatures ranging from habitability to Darwinian evolution. This instrument would focus on potential biomolecular components and patterns of complexity as a generic indicator of life.

This instrument concept can be used to measure the complexity of chemical mixture and relate it to inorganic sources, abiotic organic sources (primordial and prebiotic molecules), and organic biological sources.

Molecular Complementarity and Complexity as a Biosignature

"Complexity" has long been used to describe natural and synthetic structures that are metabolites or of biological importance[2,3]. Complexity generally refers to molecular features such as branching, cyclicity, and heteroatoms. Identifying suites of molecules that exhibit these properties can indicate biological origin.

This instrument concept does not focus on the complexity of any particular molecules but rather uses binding patterns ("molecular complementarity") of suites of molecules to generate a metric of sample complexity.

Unlike an antibody ligation assay that seeks out a particular molecule or "biomarker" as an indication of life or a specific metabolism this method uses libraries of nucleic acids capable of a **wide array** of geometries and binding patterns to **assay the complexity** of a chemical or chemical surface.

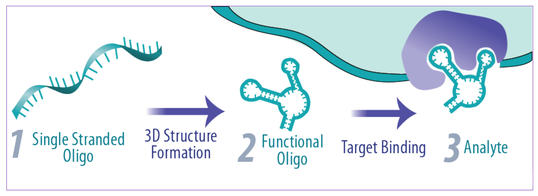
Our overarching goal is to distinguish samples with chemistries suggestive of biology by reading patterns **('fingerprints")** of molecules arising from the vast amount of chemical information on the surface of a primitives cell.

References: 1) Conrad, P.G. & K.H. Nealson (2001) A non-Earthcentric approach to life detection *Astrobiology,* 1(1):15-24. 2) Nicolaou, K.C., Hale, C.H.R., Nilewski, C. & H.A. Ioannidou (2012). Constructing Molecular Complexity and Diversity: Total Synthesis of Natural Products of Biological and Medicinal Importance. *Chem Soc Rev.* 41(15): 5185-5238. 3) Ertl P. & A. Schuenhauer (2009). Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J Cheminform.* 1:11. 4) Sun H. & Y. Zu (2015) A highlight of recent advances in aptamer technology and its application. *Molecules,* 20, 11959-11980. 5) Cleaves, H.J., Crapster-Pregont, E., Jonsson, C.M., Jonsson, C.L., Sverjensky, D.A., & R.A. Hazen (2011) The adsorption of short single-stranded DNA oligomers to mineral surfaces. *Chemosphere,* 83, 1560-1567. 6) Pai, S.S. & A.D. Ellington (2009) Using RNA aptamers and the proximity ligation assay for the detection of cell surface antigens. in. *Biosensors and Biodetection* 385-398. 7) Zamora-Olivares D., Kaoud T.S., Jose J., Ellington, A., Dalby, K.N. & E.V. Anslyn (2014) Differential sensing of MAP kinases using SOX-Peptides. *Ang. Chemie,* 126, 14288-14292. 8) Goodwin, S., Gade, A.M., Byrom, M., Herrera B., Spears, C., Anslyn, E.V. & A.D. Ellington (2015) Next-generation sequencing as input for chemometrics in differential sensing routines. *Ang. Chemie,* 127, 6437-6440. 9) Johnson, S.S., Anslyn, E.V., Graham, H.V., Mahaffy, P. & A.D. Ellington (2017) Fingerprinting non-Terran biosignatures *Astrobiology,* submitted. Castro-Wallace, S.L., Chiu, C.Y., John, K.K., Stahl, S.E., Rubins, K.H., McIntyre, A.B.R., Dworkin, J.P. Lupisella, M.L., Smith D.J., Botkin, D.J., Stephenson, T.A., Juul, S., Turner, D.J., Izquierdo, F., Federman, S., Stryke, D., Somasekar, S., Alexander, N., Yu, G., Mason, C.E. & A.S. Burton (2016) Nanopore DNA sequencing and genome assmebly on the ISS.

An Instrument Concept using Aptamer Binding to Quantify Molecular Complexity

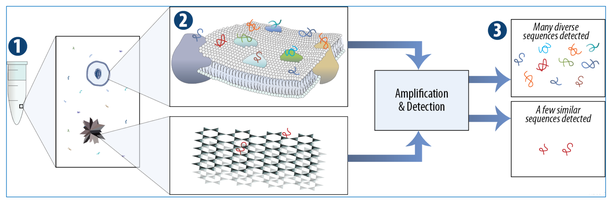
Patterns of binding to nucleic acids, independent of their biological function, can be used to probe and report on any chemical environment, opening up a new way to detect biosignatures.

Our concept builds on the fact that oligonucleotides (single-stranded DNA sections) naturally form second and tertiary structures that can have **extremely high affinity and specificity for binding** a variety of molecules, including metals, minerals surfaces, small organics, peptides, and proteins[4,5]. DNA sequences as short as 15 nucleotides can form complex structures (known as aptamers) that systematically bind to analytes in both solid and and solution samples, even in complex mixtures.



Analysis Pathway for this Instrument Concept includes:

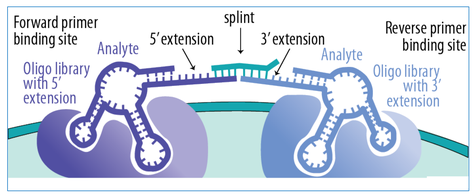
- ❑ Microfluidic sample preparation sorting by particle size or dilution to increase specificity of binding and decrease sample pool.
- ❑ Sample exposure to libraries of randomly generated aptamers that bind to molecules and particle surfaces with reproducible specificity.
- ❑ Optional *in vitro* identification of the randomly generated aptamers by the process of SELEX (Systematic Evolution of Ligands by Exponential Enrichment.
- ❑ Substrate washing to release ligated aptamers.
- ❑ Amplification of bound sequences by PCR (Polymerase Chain Reaction) enabling analysis of very low biomass samples.
- ❑ Sequencing of bound nucleic acids and/or optical detection by DNA microarray
- ❑ Statistical analysis to derive complexity patterns and relate to known biological samples



Sequencing Power: Our method uses nucleic acid binding to chemicals and on chemical surfaces to provide information about the diversity of binding sites in a sample. Many diverse folded oligonucleotides will bind to a complex surface, such as a cell membrane. Far fewer bind to simple, repeating, inorganic crystalline structures. Bound sequences can be amplified by PCR yielding high resolution data even in low-binding site samples.

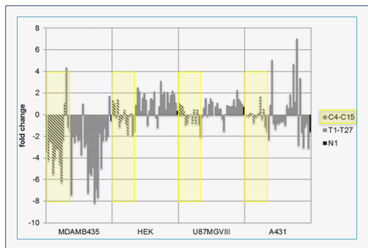
Method optimization to decrease background signal in low biomass, low-binding site samples: The Proximity Ligation Assay (PLA) is known to generate binding sequences that can be amplified and detected by sequencing or hybridization. Most importantly, this method creates very low background (i.e., non-specific DNA) of non-ligated, non-amplified species[6].

In PLA, two different DNA libraries are reacted with sample; one with a constant region on the 5' end and the other with a constant region on the 3' end. When two library members bind proximally a DNA ligase covalently joins them creating a sequence that can be amplified by primers at the constant regions. Thus, only the ligated species will be amplified. The advantage of PLA over conventional aptamer binding is that the probability of quaternary association is low so only analyte-dependent ligates are amplified. In practice, PLA show almost **no background**. This instrument would capture only high affinity binding species

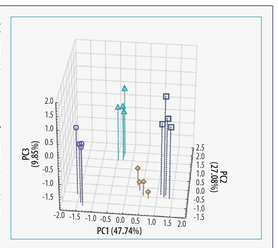


Chemometrics: Statistical Analysis to Identify Complexity Patterns Associated With Life

The sequences of bound nucleic acids can be translated into a rich information landscape. Unlike other ligands, ligated nucleic acids can be directly sequenced, revealing the complexity of the binding chemistry, regardless of the sample source. Nucleic acid strands carry information inherent in their sequence. **A sequence is analogous to a string of numbers and a sequence count is the number of times that string occurs.** By accumulating large numbers of binding sequences that reflect different compounds in a mixtures statistical analysis can identify patterns associated with increased complexity[6,7]. **This method is entirely agnostic requiring no prior knowledge of the surface attributes or 3D structures of the bound nucleic acids.**



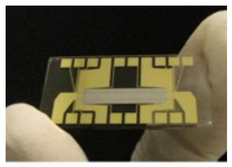
Finding Patterns in Data: The plot on the left shows **fold changes related to aptamer binding** for four cell lines. Some aptamers bind only to particular cell lines while others bind with multiple lines.
The plot on the right is the PCA score plot of all four cell lines based solely on the variance in the fold change of the aptamer pool. With no prior bias, the pool of **aptamers could pattern differences between similar cell surface attributes**. (from Goodwin *et al.,* 2015)



A number of multivariate statistical analyses can be used to generate **a fingerprint** of a sample's complexity from chemical data[9].

- ❑ PCA (Principle Component Analysis) merges data from many variables into a smaller number of axes to graphically represent the overall trends and associations. This method generates sample scores based on a simultaneous ordination of the rows and columns of the data matrix and assigns correlation coefficients for all available data.
- ❑ LDA (Linear Discriminant Analysis) is a supervised classification system that maximizes the between-group variance while minimizing within-group variance. The resulting score identified if samples belong within a training set. LDA could be trained to recognize complex chemical mixtures as well as complex individual chemicals.
- ❑ SVM (Support Vector Machines) transform non-linear correlations between variance in the data into linear representations. SVM can be used to find data patterns that correlate to the concentrations of specific components in a background mixture of unknown identities and varying composition.

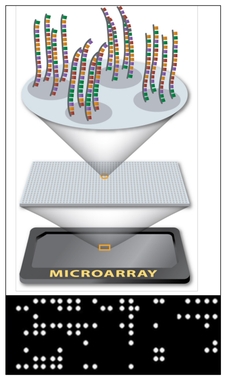
A Combination of Cutting-Edge and Deep Heritage Technologies



Our instrument relies on three microfluidic systems: sample preparation (particle size sorting solution and dilution), aptamer exposure (ligation and amplification) and detection (sequencing).

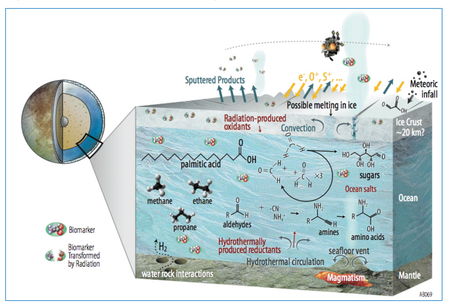
Currently we are developing a **microfluidic platform that delivers single cell sensitivity**. We are also testing commercial-off-the-shelf "**PCR on a chip**" as well as detection technologies capable of binding site pattern detection; the Oxford Nanopore's MinION Mk 1B **DNA sequencer** and a **DNA microarray**. The MinION would generate nucleic acid sequences for ligated species but would produce large data files which may limit use in the Outer Solar System. A DNA microarray could be optically or electrically imaged for a more literal "snapshot" of binding site diversity.

The MinION sequencer is small and portable (~85g), drawing less than 1W of power, but relies on characteristic changes in ionic current to identify nucleobases as a single-stranded DNA, RNA, or XNA is pulled through a protein pore. MinIONs can produce long read lengths (<100kb). Testing aboard the ISS has proven these devices to be reliable in space but further testing is required to understand how robust it is in radiation conditions typical of the Outer Solar System[10].



Once washed from the sample binding surfaces and denatured the ligated aptamers can hybridize complement sites of a DNA. A microarray the size of a microscope slide can accommodate roughly one million hybridization sites. Hybridization at a site can activate a fluorescent dye or create an electrical signal by releasing an electrochemical reporter pigment. An autonomous data analysis system could be trained to interpret microarray data.

Ocean World Applications: Our instrument concept could fingerprint intricate patterns of binding chemistry with a miniaturized device that draws little power with less than 1 µL of sample and uses an optical or electrochemical detector that requires minimal data transfer. Samples can includes particles, liquid solutions, ice or even vapor.



It is entirely likely that the amount of biomass produced lithoautotrophically on Ocean Worlds will be extremely small when compared to the biomass produced photosynthetically on Earth (**perhaps only a few tens to hundred of cells from a plume fly-by**). Yet by utilizing PCR, the signal associated with a single cell can be amplified a billion fold within hours.