Genetics and Biodiversity: Tools To Improve Geolocation and Source Attribution

Michael A. Marciano and James Crill II

Innovations in DNA sequencing, computing power, and statistical tools have paved the way for high power, DNA-based geolocation methods that are of great value to the Intelligence Community. IC analysts can now use these methods to determine where an item originated and the routes that it may have traveled. For example, DNA collected from individuals, IEDs, WMD, drugs, or computer components can be used to identify the points of origin and the travel routes of the targets. This Research Short explores how DNA-based geolocation can be used to generate probabilistic, actionable intelligence by improving source attribution to build a unique profile that provides reasonable certainty of the target’s origin and the mapping of travel routes.
DNA, found in all living organisms, has long been considered a central identifier of “who” or “what.” It can now be coupled with advanced computational and statistical approaches to answer more detailed questions about source attribution and geolocation—the process of determining the physical location(s) from which a person or object originated or was collected, or the locations that the person or object traversed. DNA can be used to overcome many challenges associated with existing geolocation methods, such as stable isotope and chemical analysis (see Appendix). This is in part due to the ease with which DNA can be transferred or deposited on surfaces. Edmund Locard’s Exchange Principle that “every contact leaves a trace” ¹ suggests biological trace materials are exchanged when there is contact between two objects or organisms. Genetic material can be deposited on items actively—when a source of biological material (human, animal, or plant) comes into direct contact with an object—or passively, such as when pollen or other particles are dispersed through the wind or “hitch-hike” from contact to contact.

Exploiting DNA-based geo-location depends on four factors, depicted in Figure 1. Many studies have explored the first factor: how to efficiently collect, isolate, and generate profiles from a biological sample. Therefore, this Short will focus on the final three: establishing a DNA sequence reference database to identify components of the biological sample; creating a catalogue of biodiversity-based range data—an organism’s spread across regions—to map identified DNA in a biological sample;* and assessing the identification statistically to lend confidence to the conclusion.

**Methods for Genomic Interrogation**

Genetic material must be isolated and prepped prior to analysis (Step 1 in Figure 1). The methods used are generally grouped into two categories: target-specific and target-agnostic. Target-specific methods, such as Polymerase Chain Reaction and amplicon sequencing, are used to target DNA sequences specific to a species. This analysis typically focuses on qualitative assessments (presence/absence of the species in a sample) and has a high level of sensitivity and specificity, but it requires a priori knowledge of the organisms of interest. In contrast, a target-agnostic approach examines all organisms in a sample, giving a global view

---

¹ Factors 2 and 3 are heavily reliant on bioinformatic analysis—the use of computational tools to capture, analyze and interpret biological data.
of the genomic community within it. Whole genome sequencing (WGS), used to identify an organism’s entire genome, provides broad information on the sample but lacks sensitivity and specificity. The approach is largely dictated by the questions that need answering.

How Can DNA Be Used To Identify Components of a Biological Sample?

DNA is a set of “instructions” written in a genetic code made up of four chemical bases—adenine (A), cytosine (C), guanine (G), and thymine (T)—where their sequence influences an organism’s developmental, metabolic, and general functionality. Many genes (functional DNA sequences) are well conserved, meaning their DNA sequence does not change over time or across similar organisms. For example, genes responsible for respiration are highly similar (in sequence) across all land animals. However, some DNA sequences vary between organisms or individuals. Any mutations, particularly in the more conserved regions of DNA, can be used for genetic discrimination of one organism or species from another (e.g., chimpanzees from humans); geographical or varietal subgroups or individuals within species populations based on geographical distribution (e.g., humans of different ethnicities or opium poppy species grown in different locations); and individuals within a species (e.g., human forensic identification).

Generally, these types of genetic identifications are only possible when a known reference sequence of the organism or individual is available. Curated reference DNA databases are critical to genetic identifications because they provide the “raw” data needed to identify DNA sequences obtained from investigative samples—Step 2 in Figure 1. For example, accurate identification of the organisms in a sample collected on a surface in a clandestine laboratory requires that the “unknown” DNA sequences be compared to known DNA sequences; without these reference DNA sequences, the organisms contained within a sample could not be identified. The most noteworthy database, the National Institute of Health’s National Center for Biotechnology Information (NCBI) GenBank database, developed in 1982, houses more than 2.1 billion DNA, RNA, and protein sequences, representing more than 478,000 species. More than 2,200 new species are added each month. Other influential curated DNA references databases include the Barcode of Life Data System (BOLD), the International Barcode of Life, PLANiTs, and the European Molecular Biology Laboratory European Bioinformatics Institute (EMBL-EBI).

Biodiversity Range Data

Step 3 requires assigning bio-geospatial range data to the identified species, again requiring reference databases—this time with the distribution of species across geospatial regions. Like the reference DNA databases, these have become more accessible in light of public interest, environmental conservation, and advancements in computing and data sharing. The Global Biodiversity Information Facility (GBIF) is the largest global biodiversity network, containing more than 1.8 billion identified species and occurrences of the species across 60,356 datasets and 5,882 peer-reviewed studies. Others are the Biological Collection Access Service (BioCASE), the Integrated Digitized Biocollections (iDigBio), the Global Genome Biodiversity Network
(GBBN),\textsuperscript{12} and iNaturalist.\textsuperscript{13} All provide high resolution bio-geospatial data that can be used with DNA-based data to isolate the likely origin of a DNA sequence.

The value of the data (DNA sequence and bio-geospatial) depends on the quality of the DNA sequence identification and the accuracy of the range data. Figure 2 shows GBIF-derived bio-geospatial range data for six organisms. The rock dove, observed worldwide, provides little value as a taxonomic indicator in determining the origin or travel routes of a target. The more specific geospatial distributions of the other species make them better taxonomic indicators—organisms with well-defined bio-geospatial distribution and DNA sequences. Identifying these taxonomic indicators in an unknown sample will help identify an item’s travel route and origin.

\textbf{Figure 2.} Example of GBIF distribution data.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{gbif_distribution.png}
\end{figure}

\textbf{Statistical Assessment: Predicting the Point of Origin and Travel Routes}

Although the corresponding DNA sequence and bio-geospatial range data are central to the mapping process, that data alone does not lend weight to an identification of point of origin or travel route. The data’s value requires Step 4: assigning a statistical confidence to the DNA-taxonomic identifications. This \textit{Short} proposes using McKay’s combinatorics approach to probabilistic geolocation,\textsuperscript{14} where an analyst calculates probabilities associated with an object originating and traversing specific routes only after enumerating all possible travel routes based on the bio-geospatial data; calculating how often each potential location appears in each possible travel route; and considering the soundness of user submitted information. Because this method calculates probabilities based on direct evidentiary data, not empirically derived data, it does not depend on human expertise or predictive mathematical models. Also, the output is reproducible, unlike Markov chain Monte Carlo methods that draw samples randomly to approximate the probability distribution of attributes.
over a range of objects, and that, although accurate, are time-consuming, computationally expensive, and not reproducible.

**Hurdles Remain**

Applying DNA analyses to law enforcement and national security issues has strengthened investigative and prosecutorial approaches. Agencies whose primary mission is to eliminate security threats increasingly rely on biological samples to aid in the identification of humans and infectious agents and to combat trafficking. Law enforcement and intelligence agencies can navigate challenges posed by resource deficiencies, technology transition, and public scrutiny by collaborating with research scientists with expertise in exploiting DNA-based investigative intelligence. Understanding these challenges requires a look into the anatomy of a forensic sample and how databases can be used to exploit the information content.

**Capturing Often Discarded Data from Forensic Samples**

Using environmental samples to geolocate has not been commonplace due to the genetic complexities of forensic samples consisting of many organisms and containing environmental components—dust, soil, water, and fomites (clothes, air filters, etc. that may carry infection)—that can make identification challenging. Typically, a forensic sample is scrutinized for markers that will identify a single species of interest or a specific person or group of people; the auxiliary genetic information (other species present) is then discarded. The discarded data, however, is critical to this Short’s proposed geolocation and source attribution method because they can be very useful in identifying where an item of interest originated or traversed.

**Challenges**

DNA-based geolocation depends on comparing collected DNA sequences to samples in genomic databases, a process challenged by poor curation, data management, and other hurdles.

**Database curation.** A database only provides value if the data is accurate, but many lack consistent curation and validated data. Bernasconi says one of the most challenging aspects of a multisource database is determining if the data is up-to-date, accurate, and trustworthy, and he argues that work in data integration will require more “data quality-aware” operational approaches. The ideal genetic database would be all-inclusive and contain the genomic sequences for native and cultured species that have a clear provenance. Most, however, are curated for a single purpose, such as foodborne pathogens, which limits their utility outside of the intended use. A geolocation database will require a combined set of data from the three primary domains of life—bacteria, archaea, and eukaryote. As of 2021 no combined database exists that contains geographical data, functional traits, phylogenies, and genetic identifiers for these domains; however, the data to build such a database does exist. Organizations and institutions such as AgBioData, GrainGenes, the Genomic Observatories MetaDatabase, and Syracuse University are laying the groundwork for standardizing how data is represented.
and for responsible sharing of data. Their models will be pivotal in creating and maintaining reliable and usable attribution databases. In addition, there are gaps in the existing databases; for example, biodiversity data is lacking from countries such as China and Russia. These databases can thus be improved through sampling from mission-specific locations of interest.

**Data Management.** Data “cleaning” and merging are critical in big data projects, yet often overlooked. This can undercut the quality of analysis and delay development. Incorporating GBIF data with ArcGIS mapping software raises challenges ranging from standardizing the number of characters in a column header to more substantial issues of inconsistent or missing data across entire sets; even the presence of white space can cause computational issues. Addressing this from a data science perspective is key to implementing a DNA-based geolocation system. Merging data collected from multiple sources will also remain a challenge unless cataloguing is standardized or methods are developed to appropriately address this.

**Chasing the Microbial “White Rabbit.”** Having the full genomic sequence is the most reliable way to identify novel and geo-specific biodiversity and is important when using microorganisms for geolocation. However, a 2016 archaeal and bacterial census shows most new sequence data is limited to shorter sequences, where only 56,000 16S rRNA genes have been identified among the 10 million to 1 billion bacterial species. A more recent FDA study found that only 1.1 percent of the genomic samples for salmonella in NCBI’s Pathogen Detection Database had data associated with collection location and only 0.6 percent included the medium in which the bacteria were isolated. In some cases, the available information simply needs to be compiled from a variety of sources. As genomic sequencing becomes more accessible and affordable, we can expect more complete sequencing and complete metadata.

**Trusting your Results.** Actionable data is only possible when the integrity of the data set(s) is preserved through vigorous testing and validation. Initial in-silico (computer-based experimentation, simulation, and modeling) testing is useful; however, the data sets will need to simulate the complexity of “real-world” samples. Laboratory testing will require the development of trustworthy control materials (known geographic identity and genome sequence) that can be used to validate a database. Furthermore, these controls will need to be validated in the presence of realistic backgrounds commensurate with those that will be used outside the laboratory to provide more accurate analysis, given that unknown samples may contain unexpected or unknown organisms. Reproducibility and sensitivity studies will be needed to ensure the method is suitable for field use—identifying the limits of detection and assessing the positive, negative, false positive, and false negative rates. Validation guidelines have been established for sequencing in the microbial forensics field and by the FDA for detection of pathogens in food that provide reliable criteria for acceptance.

**Shifting to Ecological Boundaries.** Effective database management must recognize that DNA-based geolocation transcends political boundaries. Ecological, geographic, and climate boundaries constrain an organism’s range and are most valuable in determining travel routes and points of origin. For example, it is possible to conclude that a species found well outside its ecological boundary would indicate it has been relocated by outside interference.
Summary and Outlook

Developing a reliable DNA-based geolocation tool that can help intelligence analysts identify an object’s source and routes traveled will require well-curated databases; this is becoming possible through the increased availability of biodiversity data and more widespread use of WGS as a method for analyzing complex metagenomic samples. Several existing databases are extensive and informative but lack the needed genomic content. Furthermore, to attribute genomic data to a specific geographic location, WGS data must be placed in context using a carefully curated global database layered with geographic coordinates, phylogeny, and genetic markers. Efforts to combine these tools will enable the benefits of WGS while providing sample provenances for complex sample types. Protecting data integrity is critical to the method. The database that underlies this method can include publicly available data, but it should not be publicly curated. Barring submissions from the general public will protect the data’s integrity, including the sample provenance. Responsible use of this method relies heavily on the initial validation and testing, when any limitations can be identified and used to qualify the conclusions with a specified level of confidence. Finally, conclusions generated from this data should be used in concert with other intelligence to add credence to hypotheses or to corroborate other pieces of intelligence.

Dr. Michael Marciano is Director for Research in Syracuse University’s Forensic and National Security Sciences Institute and a Co-Investigator in its IC Center of Academic Excellence. With degrees in forensic molecular biology and structural biology, biochemistry, and biophysics, he researches practical applications of molecular biology and genetics, emphasizing genetic identity, DNA-based geolocation, and computational means to interpret complex genetic data.

James Crill is a Professor of Practice at Syracuse University’s Forensic and National Security Institute and a Co-Investigator at its IC Center of Academic Excellence. He has 10-plus years of government contract research experience and is a subject matter expert with field experience in detecting and characterizing biological WMD for DoD.

Syracuse University’s bioforensics laboratories have generated the architecture for a unified genetic and biodiversity geo-sourcing database and are developing means to increase the value of biological samples from sensitive sites. Its diverse team of bioinformaticists, molecular biologists/forensic scientists, and statisticians use DNA-based geolocation for bio-geospatial sourcing and prediction of travel routes with high confidence because of the quality of the laboratory, computational methods, and databases.

This Short is a collaboration between Syracuse University’s Intelligence Community Center of Academic Excellence (IC CAE) program and NIU under the Office of the Director of National Intelligence to engage in research involving subjects of interest to the IC. NIU publishes this product as a courtesy and does not endorse the opinions of the authors.

If you have comments, questions, or a suggestion for a Research Short topic or article, please contact the NIU Office of Research at NIU_OOR@dodiis.mil.
Appendix

Stable isotope analysis is used to determine the relative amounts of specific stable isotopes—nonradioactive forms of an element that differ in atomic mass—in a sample (water, soil, vegetation). The isotopes vary based on the location of the sample. Thus, this analysis can help identify a target’s origin and serve as a surveillance marker; it has been used to identify sources of drinking water, soil, and vegetation, as well as to track migratory cycles. However, using stable isotope analysis is pricy; the analysis is highly complex and requires comparative databases that are under-curated*.

Chemical-based geosourcing is commonly used in the food and drug industries and in law enforcement and the intelligence communities. One example is the U.S. Drug Enforcement Administration (DEA) Heroin Signature Program (HSP) that uses chemical signature methods29, 30, 31 to track the source and purity of heroin.32 The tests identify the heroin’s chemical signatures created during the refinement process which can correspond to regional practices. Unfortunately, the complex and covert drug trade’s ever-changing refinement practices may limit the utility of chemical analyses. A more stable form of geolocation and source attribution could mitigate the gaps in chemical and isotope analyses.

* The term under-curated refers to a general lack of care in updating, verifying, and reviewing data within a database. Examples include the absence of real-time updating of metadata such as new species observations and inconsistency in the formatting or data.
Endnotes


