Continuous Dynamic Annotation of Environmental DNA Sequences

Continuous Dynamic Annotation of Environmental DNA (eDNA) Sequences

Rapidly assign and continuously improve taxonomic identifications for sequences derived from eDNA samples, or from other field samples lacking species identification, based on best available curated libraries of reference sequences.

Rationale

In the next few years, environmental DNA is likely to represent the largest single source of direct observations of species diversity and community composition across all systems and in all regions, introducing a rich picture of biodiversity patterns from taxonomic groups which are often currently sampled only poorly or not at all. DNA-based approaches will also in many cases greatly simplify access to rapid and accurate identifications, even of individual target organisms. Efficient tools will be required to handle the expected volume of sequences requiring identification, but there will also be continuous opportunities to correct or improve these identifications over time, as expert-curated libraries of reference sequences improve. As a result, a common infrastructure is required to assign an identifier to each unique field-captured sequence and to maintain a mapping between each such identifier and the current best assignment of the associated sequence and a known taxon. This mapping layer will maximize efficiency in real-time processing of sequences, while allowing the interpretation of these sequences to improve to reflect the best evidence available.

Status: **RESEARCH** Value: **VERY HIGH** Readiness: **MODERATE**

Estimated costs: *Research and Development* - **€1,000,000** *Operationalisation* - **€1,000,000** *Ongoing annual* - **€500,000**

Elements to accommodate

- Existing reference sequence libraries (e.g. BOLD Systems)
- Existing research networks capturing eDNA for different systems
- BOLD Barcode Index Numbers (BINs) and other existing schemes to cluster sequences to infer operational taxonomic units (OTUs)

Remaining challenges

- Efficient model for mapping eDNA sequences to fixed identifiers
- Model for accommodating and improving reference sequence libraries
- Mechanisms to accommodate expert annotations and corrections

Continuous Dynamic Annotation of Environmental DNA Sequences

GBIO Component	Significance of this investment
Modelling Biological Systems	Improve recognition and measurement of species co-occurrence within communities
Integrated Occurrence Data	Likely to be the largest single source of future occurrence data Coverage of taxonomic groups and systems which are currently poorly recorded Good potential for recording community composition
Sequences and Genomes	Essential foundation for efficient interpretation and processing of field-collected sequences

Supporting stakeholders

>>> Institutional logos here for stakeholders with particular interests in promoting delivery of this component <<<