Indexing for Life

Deliverable 2.5. Workshop 2 Policy Report: i4Life Environmental Genomics workshop report

Work package 2

Editors: Stephane Riviere, Guy Cochrane, Alastair Culham and Magda Sitko

Contributors: Gianluigi Cardinali, Guy Cochrane, Alastair Culham, Mehrdad Hajibabaei, Jean-Yves Rasplus, Sujeevan Ratnasingham, Stephane Riviere, Vincent Robert, David Schindel, Magda Sitko, Vincent Smith, Pelin Yilmaz

31 October 2012

Capacities Programme of Framework 7: EC e-Infrastructure Programme – Virtual Research Communities - INFRA-2010-2

Grant Agreement No:	261555
Project Co-ordinator:	Dr Alastair Culham
Project Homepage:	http://www.i4Life.eu
Duration of Project:	36 months
Start Date:	November 2010
End Date:	November 2013

i4Life Environmental Genomics workshop report

European Bioinformatics Institute, Welcome Trust Sanger, UK. 25th-26th September 2012

The i4Life Environmental Genomics workshop took place at the European Bioinformatics Institute, Hinxton, UK, on 25th afternoon and 26th morning September 2012.

List of participants:

Dr. Gianluigi Cardinali, gianlu@unipg.it
University of Perugia, Perugia, Italy

Dr. Guy Cochrane, cochrane@ebi.ac.uk
EBI-EMBL, Hinxton, United Kingdom

Dr. Alastair Culham, a.culham@reading.ac.uk
Reading University Herbarium, Reading, United Kingdom

Dr. Mehrdad Hajibabaei, mhajibab@uoguelph.ca
University of Guelph, Guelph, Ontario, Canada

Dr. Jean-Yves Rasplus, rasplus@supagro.inra.fr
INRA, Montpellier, France

Dr. Sujeevan Ratnasingham, sratnasi@uoguelph.ca
University of Guelph, Guelph, Ontario, Canada

Dr. Stephane Riviere, sriviere@ebi.ac.uk
EBI-EMBL, Hinxton, United Kingdom

Dr. Vincent Robert, v.robert@cbs.knaw.nl
CBS-KNAW Fungal Biodiversity Center, Utrecht, The Netherlands

Dr. David Schindel, schindeld@si.edu
Consortium for the Barcode of Life, Washington DC, United States of America

Dr. Magda Sitko, m.h.sitko@reading.ac.uk
Reading University Herbarium, Reading, United Kingdom

Dr. Vincent Smith, vince@vsmith.info
Natural History Museum, London, United Kingdom

Dr. Pelin Yilmaz, pyilmaz.mgx@gmail.com
Max Planck Institute for Marine Microbiology Research, Bremen, Germany

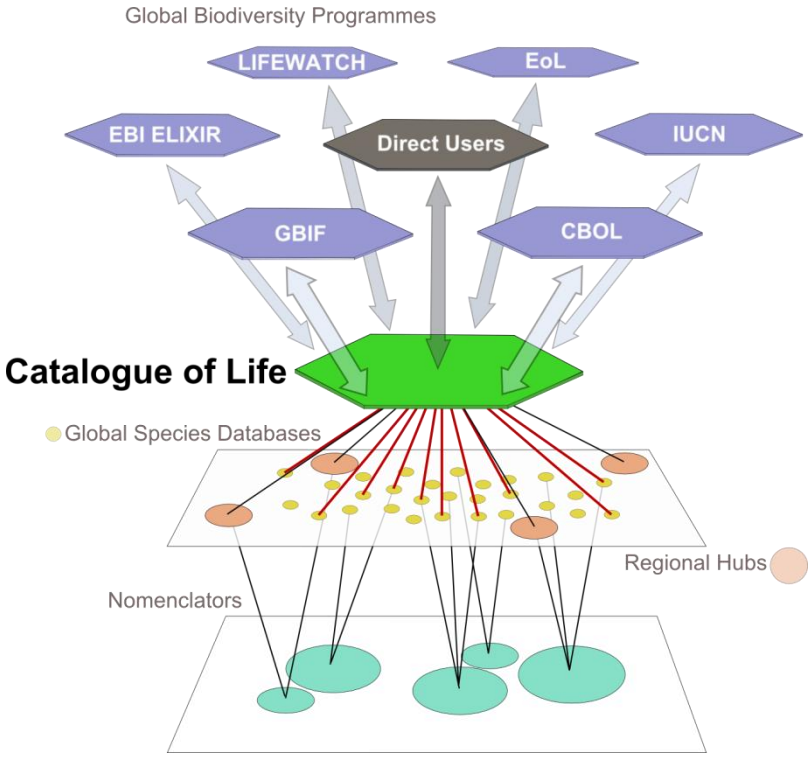
Summary of talks:

Dr. Alastair Culham: Opening and objectives of the workshops (from the 24th of October)

The barcoding and environmental genomic workshops were organised back to back as the sets of people invited for both workshops were very much overlapping. Dr. Alastair Culham gave the introductory presentation for both workshops on 24th October).

Dr. Alastair Culham (AC), the i4Life project coordinator gave a brief overview of the objectives of the i4Life project which is to establish a Virtual Research Community to interlink and harmonise global taxonomic catalogues. The existing Catalogue of Life (CoL) is used as a backbone. This builds on the work of the 4D4Life Project. AC gave a brief historic description of CoL and highlighted the tools that are available within CoL. CoL provides both dynamic and annual checklists, multilingual interface, scientific and common names, synonymy, distribution, references and scrutiny. i4Life project partners are some of the major global programmes (GBIF, EMBL-EBI, Barcode of Life, IUCN Red List, LifeWatch, Encyclopedia of Life, Sp2000, ITIS, University of Reading, ETI Bioinformatics, Cardiff University, MNHN Paris) exploring the full extent of life on Earth. This project will provide a summary of all species known across these programmes and create a global standard for taxonomic data integration in electronic infrastructures world-wide. AC described the WP of the i4Life project in details. Finally, AC gave the objectives of the current workshop:

1. To what extent are groups of sequences recognisable as taxa?
2. Can they have a stable name?
3. Should they be listed in Catalogue of Life?
4. Can i4Life begin a process of inclusion and should this be recommended?



1. Guy Cochrane: Tracking Taxa through Molecular Information.

Traditionally environmental molecular information has been generated from locus-specific studies, including 16S and 'barcoding' regions. Resolving sequences to "taxa" is possible, but there is a lack of effort to collect this information systematically. In traditional workflows, an organism name is recorded prior to sequencing. However, with environmental sequencing methods, organisms cannot be known before any data transformation following sequencing. In order to do this systematically and comprehensively we need to build a new system. In order to understand where the gaps are, we need to (i) define a very broad range of environmental sequencing applications, with examples and use cases and (ii) transform this into a set of requirements.

The system would be a Biodiversity Entry Point (portal) that may contain local data or may retrieve data programmatically. The route of access is a taxonomic question (organism with a name or with a BIN name). The system will answer the questions: Where does this organism occur? What are the associations with other organisms and environmental factors? The feed would be a taxonomic output or a set of clusters which represent potential taxa, and then the System starts beyond that.

The proposed system would be in 2 parts: a data resource (collation of information of instances of a taxon, method used for the grouping, measure of confidence…) and a service to access and search against existing taxa, mine data - at different ranks- as well as adding new taxa. The system could be centralised or a very lightweight distributed model.

Environmental data will be atomised and standardized following i.e. the minimum information, modular family of standards of MixS

(http://www.nature.com/nbt/journal/v29/n5/fig_tab/nbt.1823_F1.html).

Challenges:

- Capturing the identification methodology information.
- Simplifying the presentation of data views.
- The System has to be dynamic and respond to new methods.
- Technical.
- Organisational when we need to bring the people together.

It is important to:

- share this information early if we want people to benefit from it and accelerate the process of understanding environmental life (reverse taxonomy for example).
- understand what the gaps are (global and local). What are the priorities for taxonomy?
- calculate Biodiversity Indices such as *alpha* or *beta* diversity and what is the correlation between these taxa and environmental conditions.
- build an analytical workflow, the relationship between organisms and their environment

Regarding the procedure, we need to define some areas where the system will be important for and define use case within those areas.

Discussion: see the Global Biodiversity Informatics Conference (http://www.gbic2012.org/) November 2012 publication and the areas related to CBOL (food, endangered species, water quality…).

The objectives of the environmental genomics workshop:

1. Is it important to keep track of potential new taxa known only in sequence space?
2. Should sequence-based observation/occurrence data be collected systematically?
3. What kind of access should be required to these data; what questions will be asked of them?
4. To what extent can 'taxa' from environmental sequence studies be treated in way that is integrated across sequence-based methods and can interoperate with conventional taxonomies?

2. Pelin Yelmaz: what to do with Environmental taxa

The SILVA database currently stores more than 3 million rRNA small subunit sequences (SSU). It covers all the kingdoms: Bacteria, Archaea, Eukaryota. It contains small and large subunits for both Prokaryotes and Eukaryotes (16S, 18S, 23S and 28S). There is a bi-annual release. SILVA follows a Quality Management process with sequence & alignment quality and chimers analysis. It integrates third-party metadata (e.g. for taxonomy, nomenclature, culture and type of strains) and uses ARB for alignments and tree building (SINA alignment algorithm).

SILVA provides the alignments (through the SINA aligner) to pipelines of data analysis such as MG-RAST (WGS) or QIIME (Amplicon) and also the taxonomic classification.

The higher level taxonomic classification of SILVA is based on the 16S RNA phylogenetic trees. For cultivated species: Bacterial and Archaeal classification are based on Bergey's taxonomic outlines – not official but accepted + LPSN (http://www.bacterio.cict.fr/) - which is only valid to the class level.

3 rRNA databases in the world: SILVA, Greengenes (16S and 18S) and RDP-II (only 16S). The 3 databases have different classifications (genus level classifications are 50% different between SILVA, Green Genes, RDP). There is disagreement between databases regarding classification as the description of environmental clades is not standardised. The clades disagreement increase when we go lower in the rank. The difference between the 3 databases is not due to the difference in the methodologies but because of the environment clades names differences (from phylum to genus level). The 3 databases use different method to assign taxonomy to any sequence. RDP-II = Bayesian classifier. SILVA: 16S phylogenetic tree. Greengenes: BLAST-based approach.

There is planned roadmap of unified taxonomies: especially for environmental taxa, proposition of a cross-classification dictionary of names and alignment across the 3 databases, phylogenetic analysis to register new cluster consisting only of sequences from uncultivated organisms.

Discussion on the limitations of using only one marker: We need to think about the aggregation of more than one marker. We need to understand the limitations of using only one system. For example, for Eukaryotes, Barcoding people use 18S but also 28S and ITS. A trade-off could be having a single marker scaffold and then building off a bit as long as we support the building-off for particular purposes to have better flexibility (i.e. infectious diseases). For Fungi, beta-tubulin is used with ITS and they are moving towards 3 to 10 markers.

Disussion on methods: Amplicon and WGS (http://www.ncbi.nlm.nih.gov/genbank/wgs) are true methods, however there are methods which allow to do the biodiversity/taxonomic analysis without filtering. These are methods to analyse directly against shotgun reads. Existence of big databases using BOWTIE alignment algorithm for aligning short DNA sequence reads to large genomes and Galaxy-based Search Algorithm.

3. Vincent Smith: Making your data work for you

'Scratchpads' is a system to help share and manage Biodiversity data on the web. It is helpful for communities in particular. There are three requisites for using Scratchpads: the data must be digital, openly accessible and linkable. It allows uploading data to be published and reviewed: fast, intuitive, fit for use. Around 450 communities are using Scrachpads. An "active user" writes into the system, other registered users do not.

People consume it in different ways and there are different analytics. Scratchpads communities of one single user are the most prolific in terms of content production and those sites have the most views (trusted sets of data in their particular field, generating a lot of content and widely recognised for this particular activity).

Scratchpads version 2 was launched in March 2012. Encyclopedia of Life uses Scratchpads to build content, some museums and collections too. In order to incentivise people to push the content of their data to GBIF and EoL, Scratchpads is plugged-in into these aggregators. New forms of publication system have been explored: how do we make better use of the content that people put into those sites? Semi-automation of the production of papers to Pensoft publishers to have all the articles peer-reviewed. When it is published, Pensoft push the data to the big data aggregators ie GBIF, EoL, Zoobank. A big advantage is that not only is there a static citable version of the content, but also a light version where people can add new content. To go further, for mobilising these amount of data on behalf of the users, creation of a new concept of a general article: *Biodiversity Data Journal*.

What does it mean for barcoding, genomic and environmental sequence papers? One application is to deal with the Dark Taxa problem. Loads of data are describing clusters of sequences. They could automate the production and description of the properties of those clusters in an automated way.

For taxonomic community, we need verifiability. Taxonomic statements should be verifiable. Literature is the evidence base for taxonomy. So it is for dark taxa.

Discussion on how literature makes a description verifiable: it does in the sense that we need a citable object that we can refer to, in order to be able to say: "as best as we get it, this is the truth". Or "there is a process by which the developer of that concept went through and was published with review and this is the object we point to".

Discussion on the possibility to mine Scratchpads data: it is not possible to mine Scratchpads data at the moment but it will be possible to do so soon. There is a lot of structure underlying each of these communities so the next step will be to pull all the content together.

Plan to assign DOIs to these datasets.

Scratchpads is a good uniform way to look at Biodiversity information across different groups. As part of the portal users can use CoL classification as a backbone.

Investigation of what a data template could be for Dark Taxa. Same thing for Environmental Sequencing: what are the key bits of information that we should be enquiring from these papers?

4. David Schindel: Barcoding: CBOL's perspective

If we do have the relationship of lots of distributed data in small packets, not easy to access, not drawing a lot of attention, not as economical (Barcodes) as the big packages (Metagenomic data) then the following provocations seem relevant:

1. Taxonomic names and digitized records of collections (ie voucher specimens in Museums or repositories) are not data. They are the addresses where data can be retrieved.

2. When we get to the address, the fitness for use of the real data is not inherent in being Big data (pyrosequencing) or long tail data (Sanger sequencing).

3. Is metagenomics without a Rosetta stone (identification system for 16S or barcode reference library) a real meaningful and useful taxonomic classification or is it just ecological clustering (grouping things that occur at a point in space and time)?

There are 2 frontiers on which there is a real demand for this kind of data:

A. *Global Biodiversity Informatics Outlook* is a publication under development. It is an outgrowth the July 2012 GBIC conference. By opposition with the e-Biosphere conference (where data were taxonomic name and their geographic occurrence) the definition of Biodiversity Informatics for GBIC effort is much broader (integration from DNA up to ecosystem (VERTICAL WAY OF CHALLENGE): from DNA up to ecosystems, even remote-sensing, which is a very large view of biological data. The result is that Biodiversity Informatics needs more than just occurrence data. The new roadmap elements are represented by an ambitious set of developmental areas for Biodiversity Informatics and Environmental Genomics can be implicated in a number of potential uses:

- 1. we need more than occurrence data but traits databases and species traits. All the things that can be captured in a taxonomic description (traits) need to be turned into a computable form. So when we get to the address of the specimen or the address of the name, we just don't get when/where it was found but we get more information (colour, way of growing etc…)

- 2: intelligent multi-scalar monitoring: if we find something changing on one scale, we should investigate on adjacent scale too. For example remote sensing changes can be translated into observations seen all the way down to changes in microbial communities.

-3: expanded data horizons: going from DNA to remote sensing

-4: Genomic Observatories: monitoring things at the level of environmental genomics (Nature Reserves on another scale).

-5: Multi-scalar integration: how to connect records at the Microbiome level to the taxonomic level, and the taxonomic level into ecological communities.

-6: handling unstructured data which are not in database form (ie literature, digital images, audio recordings, video recordings). How to represent this information in computable ways or at least in locatable ways?

-7: data capture workflows.

-8: Experts' curation: how to set up a reward system for people to curate data?

-9: predictive modelling

-10: integrated citizen science

-11: Data visualisation and access

Regarding the fitness for the use of the data, the multi-scaler reach of GBIC expands to big aggregators and data standards (from DNA and genes sequences up to geographical landscape data, ie Dataone which is a metadata repository of ecological communities).

B. The second frontier is an organisation dealing with objects for scientific study. It focuses on 4 types of research: Global Change, Food Security, Zoonic and Human Diseases and Human Migration.

Scientific Collections International (http://www.scicoll.org/) covers the full spectrum of scientific collections: biomedical, anthropological, ice, sediment and rocks. SCICOLL is connects different objects based collections making them more accessible and inter-operable (HORIZONTAL WAY OF CHALLENGE). The report "Scientific Collections" provides examples of the importance and impact of scientific collections of all types based on economics, public health, innovation, public safety, or how real collection examples make a difference. The primary goal of SCICOLL is to promote interdisciplinary research. The challenge is to bring different types of data from different types of collections to bear on these problems:

- Global Environment Change (2 main sources: ice cores, providing the level of CO2 overtime or deep-sea sediment core provided the data for sea-surface temperature).
- Food Security: seed banks, gene banks, insects' collections. Ancient DNA is very important to indicate plant remains in sub-fossil soils (paleo-agriculture, crop wild relatives).

5. Sujeevan Ratnasingham: The example of Barcode Index Numbers from BOLD

The breakdown of all records in BOLD shows that: 63% of BOLD records have Linnaean names, 20% have higher taxonomy to genus or higher (result of initial sorting after collection) and 17% have interim names (based on character matching). The breakdown on the names only shows that: 29% are higher level taxonomy and only 5% interim names – this does not include BINS. The interim names part is just a place holder where people can put their names and publish it. This has been fairly consistent over the last couple of years.

Regarding the names of clusters (names of BINS), the convention is an alphanumeric standardized in size. BINS also have a DOI.

Intersection of BINS and Names: comparison of BINS of USA and Australia. For example an Arthropod has 3 different names and only one name is valid (in USA). The Australian specimen was only identified to genus (introduction from the west to Australia as a Biocontrol). Even without the names, BINS in space and time allow a non-expert, even a non-biologist to investigate further.

BINS and beta-diversity based purely on sequences: a typical molecular analysis with bacteria is represented by multi samples collection across different places without identification. By using just BINS, it is possible to get measures of the gene flow, all without taxonomic information. These data with interim names or BIN names are re-usable, so someone else could sample the same sites and re-test these hypotheses. With these data, it was possible to calculate Sorensen similarity Index for estimating beta-diversity across different samples.

BINS are beneficial in terms of costs benefits: an experiment showed that from 2 different sampling sites where malaise traps where used and everything caught sampled, there is no sign of plateau (x axis: specimen, y axis: accumulated BIN). However there is a point when this approach stops paying off: the nice thing using BINS is we know when we reached this plateau and then we move to a new site and so on (number of accumulated BINS per specimen start to plateau). The plateau means that we start to collect the same specimens.

BOLD has become an effective way of characterising Biodiversity by building the collections.

Discussion:

The workshop has set the ground for further discussions with more key players in the area of environmental genomics. We therefore plan to coordinate a White Paper to open the discussion on ways of classifying environmental genomic data and so to persuade the environmental genomics community that there is much to be learned from tried and tested classification systems. The group

attending the meeting intend to publish their discussions as an opinion paper in an appropriate refereed journal and will try to involve some additional stakeholders that were unable to attend the meeting. Community consensus is highly important in this venture.

There are two direct recommendations to CoL from this workshop:

1. That CoL recognises the value of environmental sequencing output as a complementary source of taxonomic and biodiversity information and introduce mechanisms to include environmental "taxa" within the Catalogue.
2. This workshop recommends that CoL update its management hierarchy (its high-level taxonomic backbone) in order for these environmental "taxa" to be inserted at the correct points in CoL .

Other collaborative projects, some funded by the EU include Lifewatch, BioVEL, Q-BOL and Atlas of Living Australia, from which inspiration and enthusiasm can be drawn.

Proposed possible use cases for the Catalogue of Life and genomics include: medical informatics and digitisation of medical records; gap analysis of whole genome sequences in the tree of life; sequence quality/data integrity check: see network of cryocollections and the Global Genome Biodiversity Network (http://ggbn.org/).