EMBL Explore podcast November 2012: Bioinformatics and biodiversity

Tom Livermore: Hello, my name is Tom Livermore, and in this edition you'll hear about the role bioinformatics is playing in identifying and preserving our planet's biodiversity. I spoke to Stephane Riviere and Sarah Hunter both from EMBL European Bioinformatics Institute, the EBI, in Cambridge to find out more.

Tom Livermore: The study of biodiversity conjures images of scientists surveying the plant life on remote mountains or perhaps perched high in the rainforest canopy waiting to glimpse a rare bird. But with so many species on earth, this approach alone can only get us so far. Scientists at the EBI are part of a global partnership to create a so-called “catalogue of life”, a comprehensive list where all species known to science are documented and named. But, assigning a name to all known species, a process known as taxonomy, is not as simple as it sounds. There are already 1.3 million named species in the catalogue, and DNA, or nucleotide sequences, generated daily, add to the ever-expanding volume of data to be included. Stephane, who works on the European Nucleotide Archive, the ENA, speaking from the EBI, explains more.

Stephane Riviere: The catalogue of life is an index of all living species and it currently includes around 70% of all the named species on earth. Among the partners ENA plays an important role. The ENA harbours data from a large range of organisms and each of approximately 300,000 species names is associated with at least one nucleotide sequence record and to help ENA apply names consistently and to provide the most useful indexes, ENA collaborates with the catalogue of life and a crossmapping system is also being established to relate and to map the two taxonomic systems.

Tom Livermore: The crossmapping tool is a computer programme, written by Stephane and his colleagues, which compares all the named species represented in the ENA with those present in the catalogue of life. Where there are differences, these species names are automatically sent to the global species databases – or GSDs – a consortium of over 3000 taxonomic experts around the world.

Stephane Riviere: The GSDs will review them, annotate them and then they can add them to the catalogue of life or return them to the provider, for example ENA, with notes telling us what name we should be using instead. This ensures that our data can be accessed by the broadest group of users.

Tom Livermore: This is how biodiversity informatics is helping make the vast knowledge we already have about life on earth more accessible, but it is also aiding discovery of previously unknown organisms, as Sarah Hunter, Team Leader at the EBI, explains.

Sarah Hunter: So, until quite recently the only way you could study microbes was by growing them in the lab in culture, and the problem with that is that we think that only a small number of the total number of microbes that exist on the planet, maybe about 10%, can be grown that way. One of the benefits of the new DNA sequencing technology is that you can
just sequence DNA without having to grow the organisms that the DNA comes from first.

**Tom Livermore:** This field, known as metagenomics, allows scientists to sequence the DNA of all the microorganisms present in an environmental sample such as a rainforest soil.

**Sarah Hunter:** So, there is a very, very wide range of metagenomics experiments being done just to measure how diverse the life present in different environments is. Things from what bacteria you find in people’s mouths to see how it affects their dental hygiene, the diversity of microbial life in the ocean, desert, arctic, in other habitats. My favourite metagenomics study is where someone drove around in car in the US and figured out what insects had hit their windshield.

**Tom Livermore:** Once the species within a particular sample have been identified, the diversity uncovered can reveal other important information about the function of the microorganisms and the proteins that they make. Sequencing the DNA from certain extreme environments, can reveal new kinds of organisms and new kinds of proteins that could potentially reveal new ways to produce clean energy or clean up environmental contamination such as an oil spill. However, the potential applications of metagenomics do not end there:

**Sarah Hunter:** One of the things that people are trying to do in metagenomics is take a baseline measurement in different types of environment. And then they want to see how variations in these kind of environments affect the species there and the eventual view is that either you’d be able to model what would happen if one of the factors that influence that environment, so let’s say the temperature increase because of climate change, how will that potentially have an effect on types of species that you find there. Because biodiversity of the microorganism can tell us quite a lot about the ecosystem in general it could also maybe indicate that other things could happen elsewhere in the ecosystem, like fishing stocks could be affected and things like that.

**Tom Livermore:** Scientists from a wide range of disciplines, from clinicians to ecologists, could benefit from the huge potential of metagenomics. Sarah explains how the EBI is helping them to take full advantage of this powerful new discipline.

**Sarah Hunter:** There are two majors ways that EBI wants to help metagenomic researchers: the first is that the sequencing data is quite complex and there are very large amounts of it, so one thing we want to do is provide tools that allow the scientist to analyse the data they've sequenced and give them this information about the type of organisms that are there and what functions they might have. So, that's one thing helping scientists to make sense of their data. The other is to be a repository for these sequences and so every sequence that gets submitted to the EBI metagenomics resource gets put into the European Nucleotide Archive and preserved so that future scientist can go back and reassess the data if they want to.

**Tom Livermore:** Inclusion of metagenomics sequence data in the ENA means that new species identified can be added to the catalogue of life. This is one example of how informatics is enabling the incredible biodiversity of planet earth to be documented and understood.