Potential implications of the use of digital sequence information on genetic resources for the three objectives of the Convention

Preamble

At the last CBD COP and Nagoya Protocol MOP decisions were adopted on digital sequence information on genetic resources. Ahead of next year's meeting Parties, other Governments, indigenous peoples and local communities, and relevant organizations and stakeholders are invited to submit views and relevant information on any potential implications of the use of DSI on genetic resources for the three objectives of the Convention and for the objective of the Nagoya Protocol This short paper submitted to Defra represents a contribution of the Natural History Museum London, the Royal Botanic Garden Edinburgh and the Royal Botanic Gardens Kew to this DSI debate.

Summary

- The generation of digital sequence data is growing rapidly in terms of the number of individuals and species being sequenced and the depth of genomic coverage obtained per sample.
- Three global databases, known as the International Nucleotide Sequence Database Collaboration (INSCD), exchange information and currently mediate data on more than 200 million sequences; one of them (EMBL-EBI) is subject to more than 100 million searches a year.
- Expecting the large-scale open-access international databases to regulate the use of digital sequence data from is impractical as these databases are required by their core policies (approved by national / regional authorities) not to provide barriers to seeing the sequence data or applying conditions on their use.
- Digital sequence data in the public domain provides an extremely useful resource for biodiversity conservation and sustainable management of natural resources, and supports the implementation of the Convention. The databases form *de facto* a part of the Global Taxonomic Information System called for in COP decisions. Use of the information supports identification (e.g. Invasive Alien Species), rapid inventorying of e.g. forest biota, studies on pollinators, population structures, endemism, dispersal distance, hybridisation studies, impact of dams on riverine biodiversity, conservation of genetic diversity and many others.
- No country holds sequence data for all of its biota and species likely to be intercepted by quarantine as Alien Species, pests etc. This would still be the case if sequence data were treated as a bilateral benefit between researchers and provider country. The only way in which Parties can obtain sequence data for supporting implantation of the Convention is through freely-available global databases.
- The current mechanism for sharing DSI might be considered the equivalent of a Global Multilateral Benefit-Sharing Mechanism for information.
- Ongoing publicisation of the spirit of the Convention to users of DSI from public databases is important to promote future development of partnerships and agreements.
- Our clear view is that sharing DSI without hindrance is overwhelming beneficial.
- Greater genomic coverage is particularly relevant to ABS legislation as it involves the distribution of information pertaining to biological function that has downstream exploitation potential.
- However, any modification of the current model of use of DSI would risk limiting the non-monetary benefits currently available to Parties, and consequently the implementation of the

Convention. The financial equivalence of these benefits has not been assessed, but before any action is taken it would be helpful to make this calculation and compare it (plus the implementation costs) to the revenues that might be generated by alternative models.

Availability of Digital Sequence information

The Publicly available databases

There is a well-established international framework for submitting Digital Sequence information (DSI) and making it freely available on the internet to all. The data are made available by researchers globally, and in many cases deposition in public databases is a condition imposed by journals before a scientific paper can be published. The major databases of sequences are in the United States at the National Center for Biotechnology Information (NCBI) (GenBank), Europe at the European Molecular Biology Laboratory (EMBL-EBI), and the DNA Data Bank of Japan (DDBJ). These three have entered into a formal collaboration known as the International Nucleotide Sequence Database Collaboration (INSDC). The three databases share records, so each is making available the records submitted to the others as well as those submitted directly to itself. The INDSC collaborators have agreed the following policies (Brunak et al, 2002; Lawson & Rourke, 2016).

1. *"The INSD* [International Nucleotide Sequence Database] *has a uniform policy of free and unrestricted access to all of the data records their databases contain. Scientists worldwide can access these records to plan experiments or publish any analysis or critique. Appropriate credit is given by citing the original submission, following the practices of scientists utilizing published scientific literature.*
2. *The INSD will not attach statements to records that restrict access to the data, limit the use of the information in these records, or prohibit certain types of publications based on these records. Specifically, no use restrictions or licensing requirements will be included in any sequence data records, and no restrictions or licensing fees will be placed on the redistribution or use of the database by any party.*
3. *All database records submitted to the INSD will remain permanently accessible as part of the scientific record. Corrections of errors and update of the records by authors are welcome and erroneous records may be removed from the next database release, but all will remain permanently accessible by accession number.*
4. *Submitters are advised that the information displayed on the Web sites maintained by the INSD is fully disclosed to the public. It is the responsibility of the submitters to ascertain that they have the right to submit the data.*
5. *Beyond limited editorial control and some internal integrity checks (for example, proper use of INSD formats and translation of coding regions specified in CDS entries are verified), the quality and accuracy of the record are the responsibility of the submitting author, not of the database. The databases will work with submitters and users of the database to achieve the best quality resource possible."*

It is possible to add metadata concerning restrictions on use, but the databases do not implement such restrictions.

The number of sequences available through the INSDC is 201,663,568 as of June 2017 (https://www.ncbi.nlm.nih.gov/genbank/statistics/).

A distinct type of genetic information is 'DNA Barcodes' – single genes such as COI that provide an accurate means of identification of species. Large partnerships have been developed (Consortium for the Barcode of Life (CBOL), International Barcode of Life project (iBOL)) to make available a massive 'Barcode Library' to support identification of species and strains globally (see http://www.barcodeoflife.org/content/about/what-dna-barcoding). Barcode sequences are also held by INSDC. The Barcode of Life Database (BOLD) maintained by the University of Guelph in Ontario holds barcode sequences of the COI gene; this currently makes 1.3M records public (http://www.boldsystems.org/index.php/databases). PlutoF Biodiversity Platform

(https://plutof.ut.ee/) is a data management platform for barcode data based mostly on the Internal Transcribed Spacer (ITS) region as most suitable for the identification of fungi; it has an agreement with GenBank for deposition of gene sequences, which are consequently available to INDSC.

Because only a very few genes are used as DNA Barcodes their function is well-known; they are not known to have any economic value.

The rate of use of data from these databases is difficult to ascertain. However, the EMBL-EBI search engine ran more than 100 million jobs during 2014 (EBI Annual Report 2014 - http://www.ebi.ac.uk/sites/ebi.ac.uk/files/groups/external_relations/Documents/EMBL_EBI_Annual_Scientific_Report_2014_Low.pdf), and there is no reason to suppose that the number is falling over time. There is no information on how many of these relate to the utilisation of Genetic Resources (GR) covered by the Nagoya Protocol, but it must be assumed that a proportion will do so. A common means of using these databases is to run a Basic Local Alignment Search Tool (BLAST) search, which finds regions of local similarity between sequences on the databases. To do this it may search every record.

Public databases as a Global Multilateral Benefit-sharing Mechanism (GMBSM)

Scientific research has always operated on a multilateral benefit-sharing model (although not termed as such). No Party to the Convention has the capacity to manage information on all of its biota nor the information itself. Instead all must rely on information generated and held elsewhere. If scientific information were treated solely in a bilateral benefit-sharing manner countries would not benefit from information generated about non-endemic species, or from ex situ collections. As discussed below COP has repeatedly called for greater access to information of many types, including genetic information. The development of global information systems, as called for by COP, might be considered as a GMBSM.

Use of DSI in implementation of the CBD

Use of genetic information in Conservation

Information about genetic diversity is key to conservation and sustainable use, and digital sequence information is widely used in conservation. Some is developed as a part of conservation projects, but the availability of sequence data for comparison is a vital component. For example, Laiou et al (2013), in their trial of DNA barcoding for *in situ* identification of trees as a contribution to methodology of inventorying forest resources, used GenBank as a resource and obtained 66.7% identification success. They noted that "We also faced an important limitation because the taxonomic coverage of the public reference database is incomplete." Strategic Goal C of the Strategic Plan for Biodiversity 2010-2020 is the "Improve the status of biodiversity by safeguarding ecosystems, species and genetic diversity". Without widely available information on genetic composition this goal, and particularly Target 13, would be compromised.

A Google scholar search on DNA sequencing and "conservation management" shows over 1000 scientific papers have been published between 1 January 2016 and 1 July 2017, indicating the great use of DNA sequence information in implementing the CBD. Papers cover:

- Understanding the genetic structure of populations or species to ensure effective conservation management of genetic diversity
- Techniques for DNA-based authentication, or the diversity of organisms present in a given site, when morphological identification is difficult or simply not possible (e.g. detection of invasive species; detection of protected species; identification of species in trade; identification of morphologically cryptic species).
- Understanding migrations of organisms between different sites and how that effects conservation measures
- Effects of harvest rates on genetic diversity

- Understanding changes in genetic diversity over time and potential causal factors
- Phylogenetic diversity across landscapes to assist large scale conservation planning
- Targeted sampling for ex situ collections to support in situ action
- Pollinator conservation; for example a recent issue of the important journal *Conservation Genetics* was devoted to pollinators, focussing on a variety of issues using sequence information (https://link.springer.com/journal/10592/18/3/page/1).
- Role of genes in plant development and resilience to environmental change.

All Thematic Areas of the Convention were included. It is important to emphasise that scientific publications are not divorced from implementation of the Convention. Examples of sequence information being applied directly can be seen at RZSS Conservation (http://www.rzss.org.uk/conservation/our-projects/project-search/applied-conservation-genetics/). GenTree (http://www.gentree-h2020.eu/) is a new project to provide the European forestry sector with better knowledge, methods and tools for optimising the management and sustainable use of forest genetic resources in Europe. Tools to help biodiversity managers and policy-makers employ genetic information are being developed (e.g. ConGRESS - http://www.congressgenetics.eu/default.aspx).

The evidence base for conservation planning and implementation of the Convention would be damaged if DNA sequencing is unduly restricted by the Protocol.

The Global Taxonomy Initiative and a Global Taxonomic Information System

In 1998 the Governments of the world that recognised the CBD affirmed the existence of a 'taxonomic impediment' to its implementation (Darwin Declaration) – the insufficient availability of taxonomic expertise and information that underpins successful biodiversity conservation and sustainable use. This problem was subsequently examined and solutions proposed in various Conferences of the Parties, which agreed to make information on existing taxonomic knowledge available to countries of origin, being regularly updated and available through worldwide services (COPIII/10), encouraged governments to make availability to enhance the availability of available taxonomic information, including putting it into electronic form (COP IV/1), building capacity including through international taxonomic reference centres (COP V/9), develop a global taxonomic information system (COP VI/8), develop more accessible information services for countries on their biodiversity (COP VII/9). The INSDC partners were identified as actors supporting elements of the Global Taxonomy Information System in COP IX/22. With the increase of research generating and making use of DNA sequence data in taxonomic work, the incorporation of the public DNA databases into a global taxonomic information system (also involving actors such as Catalogue of Life and GBIF) has become a very important element. From their inception, they have operated in the manner which collections of physical specimens have been called to – by making their contents and the associated data globally available to support taxonomic and other biodiversity research.

The significance of DNA, particularly DNA barcodes, within implementation of the GTI was recognised in COP IX/22 (as part of strengthening of existing networks for regional cooperation in taxonomy and to facilitate identification of Invasive Alien Species and for agricultural border inspections). In COP X/39 the ability to generate and interpret DNA barcodes in particular was emphasised as a necessary component of taxonomic capacity. The Capacity-building strategy for the GTI (COP XI/29) Explicitly referred to the need to "produce and continue to share taxonomic tools (e.g., … online tools such as … genetic and DNA sequence - based identification tools such as barcoding)", and emphasised the need to share taxonomic information. It also included a target (Action 8) "By 2019, improve the quality and increase the quantity of records on biodiversity in historic, current and future collections and make them available through taxonomic and genetic databases to enhance resolution and increase confidence of biodiversity prediction models under different scenarios." Since 2015, individuals from a number of Parties have received training in DNA Barcoding Techniques and Methodologies in the *Global Taxonomy Initiative Training Course*

on Rapid Identification of Invasive Alien Species for Achieving Aichi Biodiversity Target 9 organised through the CBD Secretariat alien species (https://www.cbd.int/gti/training.shtml). This has involved a partnership with iBOL and the University of Guelph.

Use of DNA Sequence information is now integral to the process of taxonomy and of identification, and provides a cost-effective tool for global use. For Alien Species it provides an invaluable tool because, by definition, the species are not native to the country where they are captured and are thus less likely than local species to be known to national authorities. However, the tool is only effective if it is backed by as many sequences as possible, accessible as easily as possible; this is the system built with the publicly-accessible databases.

Digital Sequence Information and the Nagoya Protocol

As noted, DSI in public databases, like other scientific information necessary to implement all aspects of the Convention, is managed in a manner similar to that of a Global Multilateral Benefit Sharing Mechanism; the information is shared openly and any Party can obtain what it needs. The non-monetary benefits arising from access to this information have been called for over many years in COP decisions.

The option discussed as a response to COP decision 13/16 is to change this model into one which operates to an extent in a bilateral manner, with Parties holding rights to published sequence data originating from specimens accessed within their borders. Some Parties (e.g. Brazil) are already stating this right. Operation of such a system without hindering its value to the whole Convention is challenging. Certainly, if it were necessary to reach an agreement with a providing country before any sequence was accessed on a database the system would inevitably fail.

The CBD objective is fair *and equitable* sharing of benefits. In the majority of cases of use of information from the public databases, the benefits generated are effectively zero – the 100 million search jobs run annually are not generating 100 million finance-generating outputs. Putting even a very small financial penalty on reading a sequence (were it to be possible) would outweigh the benefits generated and, given the number of sequences being seen, be unduly costly both for users and to implement.

Parties see a risk of monetary benefit-sharing being avoided through the download and use of sequence data in commercial activities as an alternative to accessing genetic resources *in situ*. One response to this has been to prevent the upload of those data by scientists. Because publication of data is a prerequisite of scientific publication such restrictions will lead to scientists ceasing to work on the biota of the countries concerned – they will have been prevented from publishing their research so naturally will undertake it elsewhere. Thus the perverse result will be a loss of non-monetary benefits to the country and a lack of information being generated on its genetic resources. It will also cripple the developing global resource that is already being used by Parties (the 2016 report on EMBL-EBI included comments from users in 91 non-Eurozone countries). Such a result is diametrically opposite to the scientific endeavours in support of the Convention that Parties have been calling for in many COP decisions.

Any management of DSI under the Protocol will need to be carefully targeted and manage the risk to Convention implementation through delivery of non-monetary benefits.

References

Beargrie N & Houghton J (2016) The Value and Impact of the European Bioinformatics Institute https://beagrie.com/static/resource/EBI-impact-report.pdf

Brunak S, Danchin A, Hattori M, Nakamura H, Shinozaki K, Matise T & Preuss D (2002) Nucleotide Sequence Database Policies Science 298 (5597): 1333 (see also http://www.insdc.org/policy.html)

Laiou A, Mandolini LA, Piredda R, Bellarosa R, & Simeone MC (2013) DNA barcoding as a complementary tool for conservation and valorisation of forest resources. In: Nagy ZT, Backeljau T, De Meyer M, Jordaens K (Eds) DNA barcoding: a practical tool for fundamental and applied biodiversity research. ZooKeys 365: 197–213. doi: 10.3897/zookeys.365.5670

Lawson C & Rourke M (2016) Open Access DNA, RNA and Amino Acid Sequences: The Consequences and Solutions for the International Regulation of Access and Benefit Sharing. Griffith Law School Research Paper No. 16-12. 42pp https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2848136