European Life Sciences Infrastructures and Their Implications for Plant Genomics

transPLANT report MS6

This report has been produced as an output of a meeting held at Hinxton, UK, on 1st-2nd July 2014. The aim of the meeting was to bring together groups working on plant genomics across Europe, specifically, the existing transPLANT partners, ELIXIR nodes interested in undertaking activities in the plant genomics areas, and representatives of other Europe-wide projects in adjacent areas, namely the European Plant Phenotyping Network, the AnaEE ESFRI (Analysis and Experimentation on Ecosystems) roadmap project for ecological research, and the German Network for Bioinformatic Infrastructure (Deutsches Netzwerk für Bioinformatik-Infrastruktur).

Background

transPLANT is an Integrated Infrastructures project of the European Union, funded through European Commission Directorate General for Communications Networks, Content & Technology (DG CONNECT) as part of the 7th Framework Programme for Research and Technological Development (Framework 7). transPLANT aims to establish a scalable, pan-European research infrastructure to support genomic science in plants through the organisation and interpretation of molecular data, from relatively unprocessed, experimental sequence data through to reference annotation and interpreted models. Through a combination of networking, RTD, and service activities, transPLANT will establish a new, open-access database for plant genomics, a virtual resource built from data (and expertise) distributed throughout Europe. The project is coordinated by the European Molecular Biology Laboratory (EMBL) through it's outstation, the European Bioinformatics Institute (EBI), and involves 11 partners from 7 European countries. The project is currently in its 3rd year.

transPLANT has made significant progress towards developing data infrastructure in the plant genomics area, but the use of highly data generative experimental methods is continuing to widen and there is on-going need for further activities in this area (beyond the end data of the transPLANT project 31st August 2015). The meeting had three foci (i) informational, so that each infrastructure project could be brought up to date on the current status of the other represented projects (ii) visionary, so that the participants could pool their own ambitions into a common view of the future infrastructural needs and how they might be met (iii) strategic, so that the participants could explore what approaches wrt funding, collaboration and operation) might be undertaken to pursue these goals. This report is drawn from the discussions at that meeting, and from other consultations.

Project review: Current Status

transPLANT

The transPLANT project (http://www.transplantdb.org) is now in its third year. As with all I3 projects, transPLANT contains programs for coordination, research and technical development (RTD), and service provision. The coordination activities include community engagement and consultation, standards development, and user training. Service activities are focused on developing high performance compute and cloud environments for data analysis, and providing interactive and programmatic access to plant genomic data from diverse providers in an integrated fashion. The RTD work packages concentrate on developing the data structures, tools and analysis methods needed to support the next-generation of infrastructural services. A common theme running through many of the work packages is the importance of genomic variation (which is, owing to the development of new technologies for the sequencing of DNA, becoming increasingly cheap to assay), whose linkage to phenotypes is capable of providing major shortcuts to increasing our understanding of biological function and in the development of improved crops.

Project achievements so far include considerable efforts made in all three aspects of the coordination programme. Standards activities have focused on phenotypic data as the area of greatest need (other data types are already adequately served by their own data standards), and we have worked with groups including the European Plant Phenotyping Network and biosharing.org to develop and publicise these. The service activities have seen various routes offered to provide access to transPLANT data, including a new model for integrated search allowing users to transparently query multiple resources in an independent manner. In the RTD work packages, significant progress has been made in coordinating the consistent release of well-identified, versioned genome data; a new archive has been developed for the storage, identification and future propagation of (simple) variant data, while progress has been made towards the development of new data models for more complex data (e.g. for pan-genomes); tools for online GWAS analysis and meta-data aware search have been developed; and a range of cutting-edge data analysis tools have been benchmarked and assessed. Prioritised work in the remainder of the grant period includes further activities focused on training and on standards adoption, more powerfully integrated search, browse and data analysis activities, and the integration of the various components into a prototype system for virtual plant breeding.

ELIXIR

ELIXIR (http://www.elixir-europe.org) is an ESFRI roadmap project designed to develop an infrastructure for all life science data. The model for ELIXIR is of a central hub, coordinating the activities of nationally funded nodes, to bring together national activities into a single framework and facilitate the use of nationally developed data, tools and services in the European context. As an

ESFRI project, ELIXR is core-funded by its participating Member States and has now has now been prioritised by the European Council and ESFRI for further funding by the the European Union through Horizon 2020. Future core funding will cover the operation of the ELIXIR hub, the operation of the ELIXIR governance structure, the development of pilot projects and other activities fostering collaboration amongst the nodes, the development of a universal bioinformatics service registry, and the development of standards for data representation and service provision. The individual data, services and tools should be developed by the nodes. Currently, 11 countries and EMBL have fully signed up to the implementation phase of ELIXIR, and a further 6 countries have signed Memoranda of Understanding as a preliminary step before taking full membership. Member countries have had their Node proposals approved by the ELIXIR board, and are now negotiating Collaboration Agreements to formalise the relationship between the Hub and the Node, and to set out the services that each Node will provide to the community via ELIXIR . The ELIXIR Programme for 2014-2018 has been developed by the ELIXIR Nodes and will shortly be published.

ELIXIR nodes (and putative nodes) represented at the meeting included:

The *French* node. The French bioinformatics community is currently setting up a national infrastructure of *services* in Bioinformatics called “Institut Français de Bioinformatique” (IFB). IFB serves as a unique entry point for requests of services from the Life Science community and is in charge of coordinating and structuring the activities of the regional bioinformatics platforms. IFB is the French ELIXIR node. Plant bioinformatics is one of the domains puts forward by IFB when it applied as the French ELIXIR node. The strength of IFB in this domain lies in platforms such as URGI that focuses on domestic plants (INRA) and SouthGreen (CIRAD) that focuses on “Southern” and Mediterranean plants. These IFB platforms are involved in European and international collaborations to study plants of agricultural value.

The *Italian* node. The Italian Node is organized as a Joint Research Unit (JRU), coordinated by the National Research Council (CNR), and currently includes 12 partners, among which several Universities and High Performance Computing (HPC) providers. Interest in plant genomics was expressed by the attendance of the University of Padova, one of the 12 members of the JRU.

The *Netherlands* node. ELIXIR Netherlands, hosted by the Dutch Techcentre for Lifesciences (DTL), focuses on computing and network infrastructure, teaching and education, and data interoperability. In plant genomics we will be working on all three aspects. Specific education trajectories are already set up, and we are working on integration of plant databases in our data interoperability projects (FAIRport and the "Open Data Exchange for all" ODEX4All). DTL also works on getting the individual projects in the sector together into a community, with the goal to prevent unnecessary duplications of work through frequent communications between technical project managers.

The *Portuguese* node. ELIXIR's Portugal Node provides data, tools, standards and training in the biological domain of **woody plants**, that are sources of wood, cork, chemicals and fruits (*e.g*. olives, apples, grapes, nuts and coffee). Its goal is to build an ELIXIR framework that is an added-value to forestry and related industries as well as other woody-plant based industries, and academic research in this biological domain. Woody plants are a major natural resource in Europe, with a huge ecological impact, supporting millions of jobs across diverse industries and strongly contributing to the European GDP. ELIXIR's Portugal Node is managed by BioData.pt, Portugal's national biological information network.

The Slovene node. Over the next 5 years the Slovene ELIXIR node Elixir.si intends to prioritise areas regarding storage of experimental data, infrastructure development for advanced analysis of this data and advancing the plant functional annotation in combination with ontologies. With regard to experimental data storage, the SysMO-DB SEEK initiative together with joint data management and allowable plug-in tools has been ideated. In terms of advanced data analysis, first a high-performance system will be established, capable of running and maintaining computationally intense high-throughput data analyses (e.g. CLC Genomics, Chipster for next-generation sequencing data). Lastly, in terms of plant functional annotation (gene annotation and plant ontologies), further development of GoMapMan database is planned, such as integration with other information resources available in transPLANT or external and maintenance of the database itself.

The *UK* node. The ELIXIR UK Node's initial focus is **training**. Our multi-agency approach leverages internationally recognised UK resources and expertise from across biomedical, biosciences, environmental and computational sectors. It will be delivered in specialised centres, in courses, and through e-learning, in collaboration with other Nodes and agencies. Best practices in methods, tools and standards will be promulgated. Linking the UK sectoral communities to other ELIXIR Nodes will provide access to new technologies and approaches. Some possibility exists that the scope of the UK node will subsequently widen to include particular scientific domains. Mario Caccamo, head of TGAC (a transPLANT partner), is the technical head of the UK Elixir node.

Subsequently, preliminary discussions have also been held with representatives of the Belgian node, who were unable to attend the initial meeting, but who are also interested in the area.

German Network for Bioinformatic Infrastructure (Deutsches Netzwerk für Bioinformatik-Infrastruktur)

The German transPLANT partners HMGU and IPK are participating in a national infrastructure project called de.NBI (German Network for Bioinformatics Infrastructure). de.NBI will provide comprehensive first-class bioinformatics services to users in basic and applied life sciences research. The de.NBI program coordinates bioinformatics training and education in Germany and the cooperation of the German bioinformatics community with international bioinformatics network structures like transPLANT. The concrete work packages are (I) transparent access to germplasms and germplasm metadata, (II) bridging multiple genotypes to phenotypes and (III) improved workflows for plant gene annotation. Besides the plant activities, also bioinformatics partners organized in five units from medicine and biotechnology are involved. This project will start 2015 and will be financed for five years.

European Plant Phenotyping Network (EPPN)

Plant derived products are at the center of grand challenges posed by increasing requirements for food, feed and raw materials. Integrating approaches across all scales from molecular to field applications are necessary to develop sustainable plant production with higher yield and using limited resources. While significant progress has been made in molecular and genetic approaches in recent years, the quantitative analysis of plant phenotypes - structure and function of plant - has become the major bottleneck. Plant phenotyping is an emerging science that links genomics with plant ecophysiology and agronomy. The functional plant body (PHENOTYPE) is formed during plant growth and development from the dynamic interaction between the genetic background (GENOTYPE) and the physical world in which plants develop (ENVIRONMENT). These interactions determine plant performance and productivity measured as accumulated biomass and commercial yield and resource use efficiency. EPPN (http://www.plant-phenotyping-network.eu) offers access to 23 different plant phenotyping facilities to the user community.

AnaEE (Analysis and Experimentation on Ecosystems)

AnaEE (http://www.anaee.com) is a research infrastructure for experimental manipulation of managed and unmanaged terrestrial and aquatic ecosystems. It will strongly support scientists in their analysis, assessment and forecasting of the impact of climate and other global changes on the services that ecosystems provide to society.

AnaEE will support European scientists and policymakers to develop solutions to the challenges of food security and environmental sustainability, with the aim of stimulating the growth of a vibrant bioeconomy. AnaEE will accomplish this mission by building permanent and substantial links among researchers, science managers, policy makers, public and private sector innovators, and citizens.

The AnaEE project is preparing a Research Infrastructure for experimental, analytical and computational modelling platforms related to agricultural and ecosystems science. In particular, AnaEE is focussed in integrating field and controlled environment based experimental platforms that exploit in-situ instrumentation and analytical facilities which encompass a wide range of measurements that can contribute to model development and validation. The AnaEE project expects to incorporate national nodes where the focus of work will be on genomics studies of plant and agricultural ecosystems and therefore the development of links with other International centres and projects that are specialists in this area of science will be a key component in our integration and outreach activities.

PRACE

The mission of PRACE (Partnership for Advanced Computing in Europe; http://www.prace-ri.eu) is to enable high impact scientific discovery and engineering research and development across all disciplines to enhance European competitiveness for the benefit of society. PRACE seeks to realize this mission by offering world class computing and data management resources and services through a peer review process.

PRACE also seeks to strengthen the European users of HPC in industry through various initiatives. PRACE has a strong interest in improving energy efficiency of computing systems and reducing their environmental impact.

Conclusions and Future Developments

The European Research Area is undergoing rapid evolution, with the replacement of Framework 7 with the new Horizon 20-20 program for research and innovation, and the first ESFRI projects, such as ELIXIR, entering their implementation phases. We have set up a mailing list to keep participants connected to each other, to encourage collaboration between initiatives and centres, and to provide a vehicle for seeking further funding, where required, to facilitate further collaborative initiatives. It is planned to hold annual meetings to keep the participants in contact.

The work already accomplished by transPLANT has helped establish a framework for plant genomics infrastructure and the institutional commitment of the ELIXIR nodes has the potential to build on this and to ensure the implementation of a common, interoperable approach across all Europe. While infrastructure provision is not quite an “open sport” – only certain institutions are in a position to make a stable commitment to infrastructural provision, especially at the trans-European level, the right approach does not exclude any institute or country in a position to contribute to this effort in a reasonable way; indeed, the task is too great (both in terms of scale, but also in terms of the intellectual challenge) for any such contributions to be refused. ELIXIR, in particular, with the national Nodes coordinating activities that may be local in terms of operation but global in terms of impact, provides a mechanism for making possible the coordination of a broader group (beyond the participants in any one funded collaboration) of co-interested service providers in the provision of distributed infrastructure.