CCSIS specialist EMBnet node: AGM2011 report

Alex Patak

Molecular Biology and Genomics Unit, Institute for Health and Consumer Protection, Ispra, Italy

Central Core Sequence Information System, Molecular Biology and Genomics Unit, Institute for Health and Consumer Protection, Joint Research Centre - European Commission.

The Bioinformatics Competence Groups (BCG) of the Molecular Biology and Genomics Unit (MBG) are dedicated to the collection and organisation of Genetically Modified Organism (GMO) sequence data.

GMO sequence data are not freely available and are provided to the "European Union Reference Laboratory for GM Food and Feed" (EURL-GMFF) and to "European Food Safety Authority" (EFSA), under confidentiality agreement, by applicant Biotech companies as part of the European Union authorisation procedure for the commercialisation of GMO Food and Feed.

The BCG activity is mainly dedicated to give support to the EURL-GMFF in the scientific assessment and validation of detection methods for GM Food and Feed, as part of the EU authorisation procedure, by providing a GMO sequence database integrated with bioinformatics tools (Central Core Sequence Information System, CCSIS).

CCSIS supports policy by implementing the COMMISSION REGULATION (EC) No 641/2004 which foresees following:

- the applicant shall submit the full sequence of the insert(s), together with the base pairs of the host flanking sequences needed to establish an event-specific detection method;
- the CRL shall enter these data in a molecular database;
- by running homology searches, the CRL will thus be in a position to assess the specificity of the proposed method;

Today, the CCSIS is not only used by the EURL-GMFF, but also by the:

- GMO Unit of EFSA for the risk assessment for food and feed safety;
- scientific staff of the MBG Unit in a variety of research projects.

The management of the CCSIS can be divided in two parts: *collection of GMO DNA sequence data* and *system administration of the computing facilities*.

Collection of GMO DNA sequence data

The data stored on the CCSIS comprises a) local copies of public sequence databases, and b) in-house data.

Local copies of public sequence databases

The use of local copies of public databases, like GenBank, protects the confidentiality of the GMO sequence data during the analysis, and improves calculation speeds. Local copies are regularly updated and extended with new datasets when required. The rapidly growing amount of publicly available sequence data is a great challenge for the computing platform, especially for the local storage capacity and network resource needed to copy the data.

In-house data

The most valuable of the BCG activities is the GMO sequence data that have been collected from documents submitted to EURL-GMFF and to EFSA, and complemented with information extracted from GenBank and EMBL. The CCSIS database is a unique source of GMO sequence data world-wide, as this information in not publicly available, but is provided confidentially to the JRC by Biotech companies. The sequences provided to the BCG are verified and manually annotated with all relevant information extracted form the accompanying documents. By following international annotation rules, the sequence records are compatible with most of the available bioinformatics tools.

The new GMO Detection Methods database[1] has been released on the EURL-GMFF website. This is a publicly accessible database that con-

[1] http://gmo-crl.jrc.ec.europa.eu/gmomethods/

tains all validated GMO detection methods that the MBG Unit published in the "Compendium of validated GMO Detection Methods" (JRC Reference Report).

System Administration of the computing facilities

The CCSIS runs on a dedicated high performance computing platform. This platform is based on the clustering of several compute nodes using grid computing. Its main use is the alignment of DNA sequences. The grid-computing set-up allows many users simultaneously (high-throughput computing) to perform complex sequence alignments with and acceptable computing speed.

The Apple Workgroup Cluster has 4 Xserve G5 nodes and Xserve RAID with 4 TB of capacity. Upgrade of the hardware platform has been planned and is ongoing. This includes acquisition of new hardware, update to last Operation System versions, and a new version of the iNquiry software have been prepared on a new prototype bioinformatics cluster.

The MRS Sequence Retrieval System will be separated from the cluster and installed on an Ubuntu Linux-based Hewlett-Packard server with enough RAM memory to cope with the continuously growing local copies of public sequence databases.