

The GMOs Molecular Register: an Integrated Bioinformatic System to support detection/quantification of GMOs - (session: Database: Ontology and Integration)

D. D'Elia¹, P. Leo³, G. Scioscia³, P. Lopriore³, G. Delle Foglie¹, F. Licciulli¹, M. Millot⁴, F. Weighardt⁴, L. Bonfini⁴, R. Lorberth⁵, P. Heinze⁴, G. Van den Eede⁴, M. Attimonelli², H.-J. Buhk⁵

1) Istituto di Tecnologie Biomediche - CNR, Sezione di Bioinformatica e Genomica, Via Amendola 168/5, 70126 Bari – Italy -

2) Dipartimento di Biochimica e Biologia Molecolare, Università di Bari, Via Orabona 4, 70125 Bari – Italy -

3) Java Technology Center, IBM SEMEA Sud, via Tridente 42/14, 70125 Bari – Italy -

4) European Commission, Joint Research Centre, Institute for Health and Consumer Protection, Biotechnology and GMOs Unit, TP331, I-21020 Ispra (VA), Italy -

5) Robert Koch-Institut, Centre for Gene Technology Nordufer 20, D-13353 Berlin, Germany

In July 2000, the European Network of GMO laboratories (ENGL) has been created; its official inauguration took place in Brussels on December 4, 2003. The development of a Molecular Register (GMOs MOLREG package) has been planned. The Molecular Register will contain data on molecular characterization of GMOs approved for placing on the market in the EU and the necessary on-line tools to analyse the related sequences. This project was considered of highest priority since such a kind of Bioinformatics Instruments is extremely necessary to support the ENGL in the detection and characterisation of engineered genetic constructs.

Collaboration agreement contracts for the development of the GMOs MOLREG package have been signed between the Institute for Health and Consumer Protection of the Joint Research Centre and:

- the Section of Bioinformatics and Genomic of the ITB – Consiglio Nazionale delle Ricerche (CNR), Italy

- the Robert Koch-Institut (RKI), Zentrum Gentechnologie, Germany

The CNR lead the design and development of the GMOs MOLREG package as an easy to use web-based system consisting of a database storing GMOs data, a GMOs data submission component, a query component and a Bioinformatic Tool for running GMO's bio-sequence analysis. The design of the GMOs MOLREG database itself and the development of the web-submission component has been done by the CNR in collaboration with RKI and the JRC, while to design and develop the GMO Query component and the GMO bioinformatics analysis component the CNR established a collaborative agreement with the Life Science Team of the IBM Semea Sud, a specialized services unit of IBM in Italy.

The database has been designed and developed to include information on the following aspects: general administrative/legislative information on the registered GMO, specific information on the origin of the GMO including a detailed molecular characterisation, detailed information on available GMO detection methods and certified reference materials, screening, identification, quantification methods and information on reference literature.

The need for running bioinformatics analysis on GMOs bio-sequences required that the Registry was deeply integrated with Bioinformatics analysis tools, using a novel approach based on XML, giving end-users the possibility to analyse molecular data stored into the Repository and/or correlate such data with external public biological databases (EMBL, SWISS-PROT, PROSITE, REBASE, TRANSFAC, UNIVEC). At this purpose specific bioinformatics algorithms were required to be integrated in the GMO Repository information infrastructure; they are some of the analysis program of the EMBOSS package, besides BLAST and FASTA.

An important aspect concerning the integration between the GMOs MOLREG database and the external bioinformatic analysis tools is that a specific XML-based tools description language has been developed in order to generalize the runtime-generated web user interface, both to take tool input parameters and to show the tool results. This XML language provides a mechanism to have independence between the GMOs MOLREG database and the underlying bioinformatics analysis package used. Additionally, a web administrative console that manipulates the XML descriptors was also created: to manage analysis programs updates; to attach, as a sort of plug-in, new bioinformatics programs.

Strong security features (crypting, profiling authorization and control, certificates, etc..) were also implemented in order to protect the confidential business information flow during all the steps of the extraction/analysis process.

The system has been developed in Java by using the IBM Websphere Studio Application Developer 4.0.3

environment and deployed and successfully tested on Websphere Application Server, Oracle Application server as well as on Tomcat.

The major benefit coming from the development of such a kind of Bionformatic System is that it significantly reduces the times to perform bio-sequences retrieval and analyses through the user-friendly web query and processing system that helps and guides end-users to find the data and the analyses that best fit to their need with the implementation of a workflow logic. This has a great importance in these time-consuming operations in which the researcher has to perform data retrievals and analysis manually, in many steps. Moreover, a great flexibility is provided by this application that doesn't need, as the customer requested, programming skills to be maintained: the XML interface to the analysis tools, with the administrative console, allows to simply follow the frequent updates that affect these tools by automatically changing the XML descriptors, without any additional Java source modification.

Acknowledgement: This project has been funded within the grant n.17356-2000-12 F2SC ISP IT signed between CNR (scientific responsible M. Attimonelli) and JRC (scientific responsible G. Van den Eede). Moreover funds from Progetto MURST Cluster C03/2000, CEGB have contributed to the project.