The MicroScope platform:

from data integration to a rule-based system for massive and high-quality microbial genome annotation

Jonathan MERCIER, Alexandre RENAUX, Adrien JOSSO, Aurélie GENIN-LAJUS, E'Krame JACOBY, François LE FEVRE, Guillaume ALBINI, Claude SCARPELLI, Claudine MEDIGUE, David VALLENET

Direction des Sciences du Vivant, CEA, Institut de Génomique, Genoscope, France
CNRS-UMR8030, Evry, France
Université d'Evry Val d'Essonne, Evry, France

Abstract. The emergence of the Next Generation Sequencing (NGS) generates an incredible amount of genomes, whereas curation efforts to annotate them tend to decrease despite some community initiatives (*1*). To ease this manual process, we develop the MicroScope platform: an integrated environment for the annotation and exploration of microbial genomes (*2*). It is made of three major components:

1. a management system to store and organize biological knowledge in relational databases
2. a production system to organize and execute workflows
3. a visualization system for expert analyses and data curation through a Web interface.

Following the success of the platform, we are improving its throughput analysis to integrate an increasing number of genomes in a reasonable human time while maintaining a high quality of annotations. In this way, we initiate new methodological and technical developments on the MicroScope data management and production systems. A specific focus is given on the use of rule-based systems for the management of workflows and for the consistency evaluation of functional annotations that are performed automatically and then expertised by biologists.

The MicroScope data management system is made of several relational databases. The Prokaryotic Genome DataBase (PkGDB) gathers internal genomic data, human expertise and computational results. This central model is enriched by the integration of numerous public databases collecting different types of biological entities (e.g. genomes and genes from nucleic databanks, proteins from UniProt, metabolic data from ChEBI, Rhea, KEGG and MetaCyc). To support continuous data integration and reconciliation of these external resources, we designed the Galileo (*3*) application based on AndroMDA (*4*) and Play (*5*) frameworks. The Galileo model manages the integration of several releases of the same biological resource and ensures unicity of biological objects from different resources with the use of internal business keys based on their key properties. For instance, molecules are identified through their InChl signature, and reactions by a combination of stoichiometry information

and molecule signatures. In comparison to other initiatives for biological data integration using Semantic Web, our main purpose is the unification of entities based on their common properties to define public methods and queries. These services are provided through REST API.

The MicroScope production system orchestrates about 25 workflows, which combine various bioinformatics software. The goal of this system is to keep analyses up to date according to the integration of new genomes, the updates of public databases and new software versions. Five years ago, we adopted the jBPM (*6*) framework to design our workflows. This framework allowed us to gain synchronization, robustness, control and traceability in the execution of million of jobs on HPC clusters (*7*). To increase the throughput of our analyses and the flexibility in the decision-making process of our production system, we will progressively switch to a new API called BIRDS (BioInformatics Rules Driven System) and developed at the Genoscope. BIRDS is based on the Drools framework (*8*) and provides a common environment for business rules and resource-driven workflows to automate bioinformatics treatments. This decision process integrated in a large data management system is an original feature of BIRDS in comparison to other workflow initiatives in biology (e.g. Taverna, Galaxy).

One important goal is to ease the human interpretation of genomic data in the light of predicted functions and biological processes (e.g. metabolic pathways). We are working on an explicit representation of the biological knowledge and on algorithmic tools designed to automate the biologist reasoning within the MicroScope platform. This application, named Grools, is a rule-based expert system. It is currently under development using the Drools framework. A first level of rules is designed to predict molecular functions according to protein domain composition and organism taxonomy. These rules were extracted and translated from the UniRule resource of UniProt database (*9*). The next step is to evaluate the overall coherence of these individual functions by applying logical rules between them and integrating additional information from biological processes where these functions may occur or not in a given organism. A first implementation of such deductive reasoning has been implemented in the HERBS system through a collaborative project between INRIA and SIB institutes (*10*). This can be applied, for example, crossing growth phenotypes on defined minimal media (e.g. Biolog phenotype microarray) and functional annotations to check the consistency of corresponding catabolic pathways.

In synergy with the technological progresses in the production system, biological data integration combined with logical reasoning should improve completeness and consistency of genome knowledge in the MicroScope platform. These IT innovations will be illustrated on the poster.

Keywords: Genome annotation, Data integration, Workflow, Rule-based system, Knowledge reasoning, Curation

References

1. R. Mazumder, D. A. Natale, J. A. E. Julio, L.-S. Yeh, C. H. Wu, Community annotation in biology. *Biol Direct* **5**, 12 (2010).
2. D. Vallenet, E. Belda, A. Calteau, S. Cruveiller, S. Engelen, A. Lajus, F. Le Fevre, C. Longin, D. Mornico, D. Roche, Z. Rouy, G. Salvignol, C. Scarpelli, A. A. Thil Smith, M. Weiman, C. Medigue, MicroScope–an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data. *Nucleic Acids Res.* **41**, D636–647 (2013).
3. F. Le Fèvre, A. Josso, Galileo. galileo.genoscope.cns.fr (2014).
4. M. Bohlen *et al.*, AndroMDA. *AndroMDA*, www.andromda.org (2007).
5. G. Bort *et al.*, playframework. *Play framework*, www.playframework.com (2009).
6. T. Baeyens *et al.*, A java Business Process Management. *Java Business Process Management*, jbpm.jboss.org (2004).
7. D. Vallenet, S. Engelen, D. Mornico, S. Cruveiller, L. Fleury, A. Lajus, Z. Rouy, D. Roche, G. Salvignol, C. Scarpelli, C. Medigue, MicroScope: a platform for microbial genome annotation and comparative genomics. *Database (Oxford)* **2009**, bap021 (2009).
8. B. McWhirter *et al.*, Drools. drools.jboss.org (2001).
9. UniProt Consortium, Activities at the Universal Protein Resource (UniProt). *Nucleic acids research* **42**, D191–D198 (2014).
10. A. Viari, HERBS: A rule based system for checking the annotations of complete proteomes, Personal communication, (2014).