

GenoStar: a real-size application of ontologies in genomics

Alain Viari

– Projet Helix – INRIA-Rhône-Alpes
655, Av. de l'Europe, 38330 Montbonnot-Saint Martin

Genostar is a software platform dedicated to genomic data integration and analysis. It has been developed in a collaboration between four partners : two biotech companies, Hybrigenics (Paris) and GENOME express (Grenoble), and two research institutes, in biology, Pasteur Institute (Paris), and in computer science, INRIA Rhône-Alpes (Grenoble). GenoStar has been designed according to a specific view of the genomic world as a large network of biological entities and their relationships. For instance, a gene entity is related to its protein product(s), may be linked to homologous genes, is located on a specific chromosome which, in turn, is related to an organism and so on. The platform therefore relies on an explicit modeling of this network both in terms of entities and of associations.

The purpose of this talk is not to describe in details the different models behind GenoStar but to focus on the four following aspects :

- 1) which elements do we need to express a biological model (a "fat ontology") ? This question is related to the "meta-model" that is, in some way, the language for specifying the model. This is an important point since the "expressiveness" of a model strongly depends upon its underlying meta-model and most meta-models have been designed with computer programming rather than biological data representation in mind. GenoStar relies on an "UML-like" meta-model. I will discuss some advantages and limitations of this approach in the light of its use within the platform.
- 2) once a model has been designed, it should be instantiated with data. This can be done by using either external databases (this aspect will be left to another talk by Anne Morgat, in this workshop) or by using computational methods (e.g. to find genes on a chromosome). A key feature of GenoStar is the way this methodological knowledge is also explicitly represented (using a similar meta-model). In the platform, computational methods are explicitly described through their input and output, that are themselves entities of the data model, so that the system can easily check if a method is being appropriately used or not.
- 3) once a model has been instantiated, another important aspect is to browse thru the network of instances. In GenoStar, queries are expressed as partial networks which are searched for in the network of the knowledge base. The results of such queries are themselves partial, but instantiated, networks which can be further explored. Through this exploration process, the biologist may be brought to infer new relations between previously unrelated entities. A typical example of such an inference is the prediction of the function of a gene, using information of other genes to which it is related e.g. by sequence similarity). Example of these querying and browsing activities will be given.
- 4) finally, the last (and still open) issue of this talk is related to the question of interoperation between data models. In GenoStar each software module has its own separate data model raising the question of merging pieces of information from different knowledge sources.

GenoStar website: <http://www.genostar.org>