

Representation and integration of heterogeneous genomic data: the Panoramix project

A. Morgat¹, F. Boyer², S. Bruley¹, A. Viari¹

¹INRIA-Rhône Alpes, Grenoble; ²Université Joseph Fourier, Grenoble, France

The main theme of the Helix group at INRIA Rhône-Alpes is the modeling of biological knowledge in the context of genomics. Our goal is to provide formal and explicit representations of all biological entities involved in genome analysis (at the storage or analysis levels). This formal representation is prerequisite in order to fully exploit genomic data such as completely sequenced genomes or expression data.

To date, more than 50 completely sequenced microbial genomes have been made available to the scientific community and hundreds of sequencing projects are pending.

[<http://wit.integratedgenomics.com/GOLD/completegenomes.html>]. For these projects as well as for the former one, in-depth genome annotation is still an issue. Indeed, current annotation procedures mainly focus on the **syntactic level** of the genome. That is, they focus on feature detection (genes, regulation signals,...), but not on the relationship between these biological entities. For instance we may want to answer questions like : which genes are co-regulated? which genes encode for proteins involved in the same metabolic pathways or for components of an enzymatic complex? Moreover syntactic annotation often misses some important contextual information such as the fact that enzymes involved in the same metabolic pathways tend to be encoded by genes clustered in operons. We believe that a complete genome annotation process should take into account this contextual information (conserved chromosomal organization, molecular assemblies, metabolic pathways, regulation networks...). We refer to this annotation level as the **relational level**. Representing the entities at this level is a more difficult task and requires specific tools. The Panoramix project aims at federating a set of knowledge bases dedicated to the relational annotation of microbial genomes : Genomix, Proteix, Metabolix. A synoptic summary of the Panoramix project, showing the complementarity of these three bases, is presented in figure 1.

- Genomix

Genomix gathers information on genes and gene-products for 49 completely sequenced microbial species together with information on comparative genomics (orthology) and conserved chromosomal organization (bacterial syntenies) for all couples of chromosomes in the database.

- Proteix

Proteix describes information at the protein level (polypeptides), with a special emphasis on molecular assemblies (such as enzymatic complexes) and post-translational modifications.

- Metabolix

Metabolix gathers information on intermediary metabolism (compounds, biochemical reactions, enzyme and metabolic pathways). Intermediary metabolism can be analyzed through different perspectives (chemist, biochemist, geneticist, computer scientist,...) that we try to integrate in the underlying model.

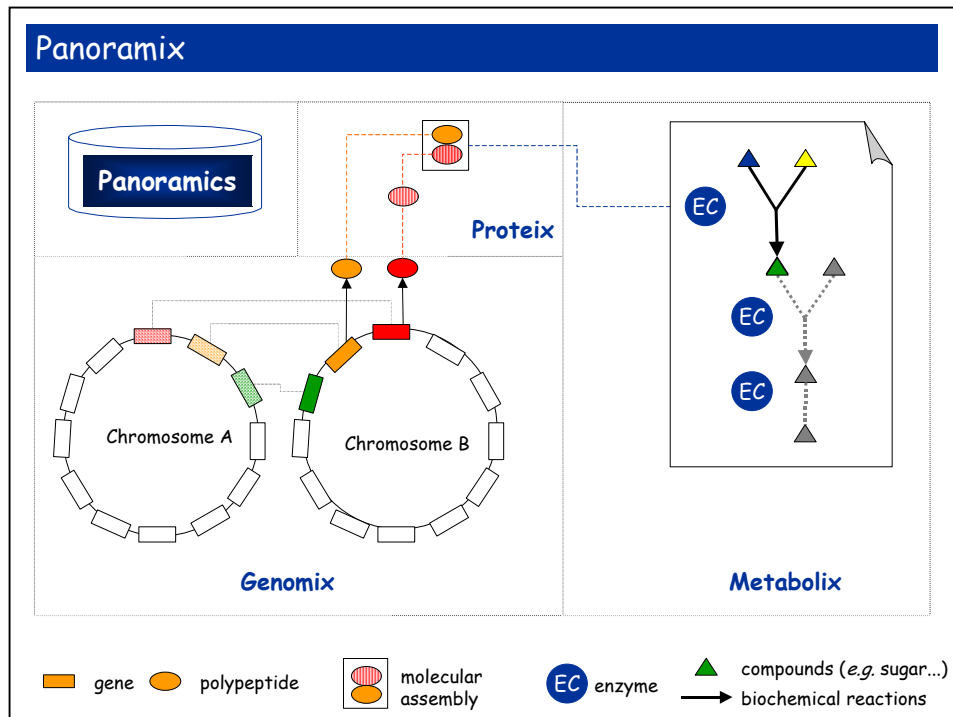


Figure 1: the Panoramix project: Integration of the Genomix, Proteix and Metabolix knowledge bases