

# The Microbiogenomics warehouse: extraction and integration of relevant information from heterogeneous data provided by genomic comparisons

J-F. Gibrat

Unité Mathématique, Informatique et Génome,  
INRA, Jouy-en-Josas

Thematic day on Integrative Genomics, Nantes, October 21, 2010

# Challenges in integrating biological data sources

- Huge amount of data stored in biological collections
- Biological data are characterized by their heterogeneity (different formats/viewpoints)
- Biological data tend to be organized around a given type of experiment
- Many questions in biology need to be addressed by combining data from multiple sources
- Database integration problem has been recognized for many years : “achieving coordination and interoperability among genome DBs and other informatics systems must be of the highest priority... We must think of the computational infrastructure of genome research as a federated information infrastructure of interlocking pieces...”
- Advances in biology are hindered not by lack of data but by the diversity of technologies used to store the data

# Challenges in integrating biological data sources

- Huge amount of data stored in biological collections
- Biological data are characterized by their heterogeneity (different formats/viewpoints)
- Biological data tend to be organized around a given type of experiment
- Many questions in biology need to be addressed by combining data from multiple sources
- Database integration problem has been recognized for many years : “achieving coordination and interoperability among genome DBs and other informatics systems must be of the highest priority... We must think of the computational infrastructure of genome research as a federated information infrastructure of interlocking pieces...”
- Advances in biology are hindered not by lack of data but by the diversity of technologies used to store the data

# Challenges in integrating biological data sources

- Huge amount of data stored in biological collections
- Biological data are characterized by their heterogeneity (different formats/viewpoints)
- Biological data tend to be organized around a given type of experiment
- Many questions in biology need to be addressed by combining data from multiple sources
- Database integration problem has been recognized for many years : “achieving coordination and interoperability among genome DBs and other informatics systems must be of the highest priority... We must think of the computational infrastructure of genome research as a federated information infrastructure of interlocking pieces...”
- Advances in biology are hindered not by lack of data but by the diversity of technologies used to store the data

# Challenges in integrating biological data sources

- Huge amount of data stored in biological collections
- Biological data are characterized by their heterogeneity (different formats/viewpoints)
- Biological data tend to be organized around a given type of experiment
- Many questions in biology need to be addressed by combining data from multiple sources
- Database integration problem has been recognized for many years : “achieving coordination and interoperability among genome DBs and other informatics systems must be of the highest priority... We must think of the computational infrastructure of genome research as a federated information infrastructure of interlocking pieces...”
- **Advances in biology are hindered not by lack of data but by the diversity of technologies used to store the data**

# Data heterogeneity

- **Syntactic heterogeneity** :  
representation **formats** are different, e.g., flat files, XML format, relational format, etc.
- **Semantic heterogeneity** :
  - At the level of the general *organization of the information* (**schema**) : different viewpoint about the entities and different attributes, e.g., metabolic pathways in KEGG and UNIPROT.
  - At the level of *instances* (**data**) : a biological entity can have different attribute values, e.g., for *H. ducreyi* gene names

| UNIPROT | KEGG |
|---------|------|
| tusA    | sirA |
| tilS    | mesJ |
| oxaA    | yidC |
| glmM    | mrsA |

# Steps for building an integrated system

Objective : to make data distributed over a number of distinct, heterogeneous databases accessible via a single interface

- 1 **Data** model transformation : resolving syntactic heterogeneity
- 2 Semantic **schema**<sup>1</sup> matching : establishing correspondences among concepts with semantic overlap
- 3 **Schema** integration : creation of a global schema (from simple union of all underlying schemas to conception of a new schema)
- 4 **Data** transformation and semantic **data** matching : establishing data correspondences in all the sources

This is a critical step in practice !

---

<sup>1</sup>As in the previous slide schema means the way of organizing the information

# Steps for building an integrated system

Objective : to make data distributed over a number of distinct, heterogeneous databases accessible via a single interface

- 1 **Data** model transformation : resolving syntactic heterogeneity
- 2 Semantic **schema**<sup>1</sup> matching : establishing correspondences among concepts with semantic overlap
- 3 **Schema** integration : creation of a global schema (from simple union of all underlying schemas to conception of a new schema)
- 4 **Data** transformation and semantic **data** matching : establishing data correspondences in all the sources  
**This is a critical step in practice !**

---

<sup>1</sup>As in the previous slide schema means the way of organizing the information



# Dimensions of integration

Numerous integration solutions have been proposed during the last 20 years. Integration of databases can be classified according to 2 axes :

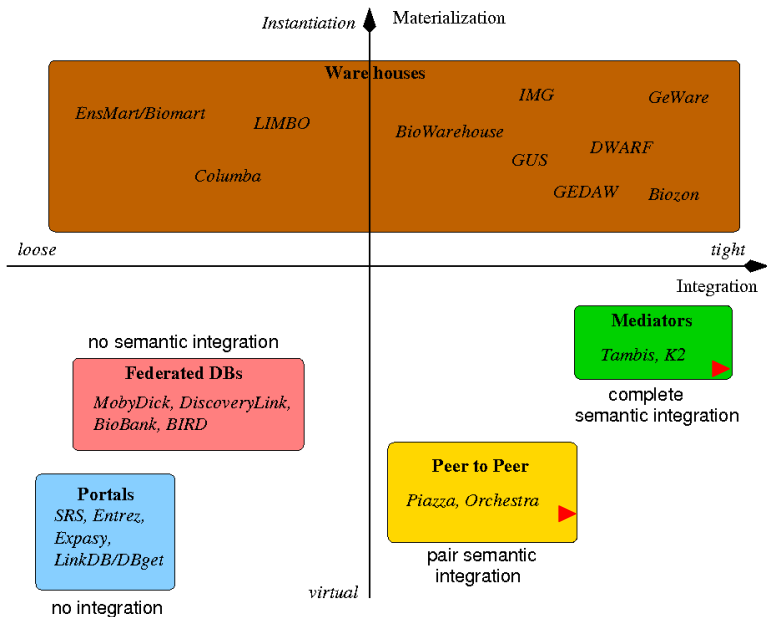
## 1 Loose vs tight integration :

- how many of the tasks of DB integration have been performed ?
- in how much detail ?

## 2 Instantiation vs virtual (view) :

- local physical copy of the DBs
- access to distant sources (e.g., using mediators)

# Examples of solutions



# Microbiogenomics warehouse

# Overview and motivations

## Microbiogenomics project

- Project funded by the French national science foundation (ANR)
- Teams involved :
  - Evolution Moléculaire et Bioinformatique des Génomes, Institut de Génétique Microbienne, Paris XI Univ., Orsay
  - Mathématique, Informatique et Génome, INRA, Jouy
  - Laboratoire Recherche en Informatique, Paris XI Univ., Orsay

## Development of a resource center for *microbial* genomics

- that gathers together pertinent data required for :
  - genome annotation/re-annotation
  - comparative genomics
  - gene/genome molecular evolution studies
- that combines heterogeneous data from different sources
- that allows the implementation of (intensive) data mining techniques

Focus on protein-related data

# Overview and motivations

## Microbiogenomics project

- Project funded by the French national science foundation (ANR)
- Teams involved :
  - Evolution Moléculaire et Bioinformatique des Génomes, Institut de Génétique Microbienne, Paris XI Univ., Orsay
  - Mathématique, Informatique et Génome, INRA, Jouy
  - Laboratoire Recherche en Informatique, Paris XI Univ., Orsay

## Development of a resource center for *microbial* genomics

- that gathers together pertinent data required for :
  - genome annotation/re-annotation
  - comparative genomics
  - gene/genome molecular evolution studies
- that combines heterogeneous data from different sources
- that allows the implementation of (intensive) data mining techniques

Focus on **protein-related** data

# Microbiogenomics warehouse : data sources

## Primary data (public collections) :

- GenomeReview/MICADO : microbial genome nucleic sequences
- UniProtKB/PROSE : protein sequences, features and annotations
- Pdb/PDB : macromolecules 3D structures
- Kegg/PAREO : metabolic/functional pathways
- NCBI Taxonomy/TAXO


## Secondary data : genomic comparisons at the protein level :

- ORIGAMI : homology relations between proteins  
comparisons of all proteins of all microbial genomes with BLAST

## Tertiary data : derived from secondary data

- orthologs and paralogs
- genomic context : synteny, phylogenetic profiles, gene fusion/fission
- GENOPAGE : protein domains/modules, protein/domain family phylogenetic trees

# Microbiogenomics warehouse : architecture

- **The warehouse is built as a relational database (postgreSQL)**
  - ▷ allows syntactic integration (entity/relation framework)
  - ▷ powerful query language (SQL)
  - ▷ can cope efficiently with large datasets
  - ▷ allows query optimizations and thus intensive data mining
- **Each source has its own schema**
  - ▷ advantage : they can be updated individually
  - ▷ advantage : it is easy to add a new source of data
  - ▷ drawback : data are not reconciled...  
but redundancy helps in discovering complementarities and divergences between sources (DB cross-checking)
- **Semantic integration**
  - ▷ semantic *data* matching : link table 
  - ▷ semantic *schema* matching : (multiple layer architecture)

# Microbiogenomics warehouse : architecture

- **The warehouse is built as a relational database (postgreSQL)**
  - ▷ allows syntactic integration (entity/relation framework)
  - ▷ powerful query language (SQL)
  - ▷ can cope efficiently with large datasets
  - ▷ allows query optimizations and thus intensive data mining
- **Each source has its own schema**
  - ▷ advantage : they can be updated individually
  - ▷ advantage : it is easy to add a new source of data
  - ▷ drawback : data are not reconciled...  
but redundancy helps in discovering complementarities and divergences between sources (DB cross-checking)
- **Semantic integration**
  - ▷ semantic *data* matching : link table
  - ▷ semantic *schema* matching : (multiple layer architecture)



# Microbiogenomics warehouse : architecture

- **The warehouse is built as a relational database (postgreSQL)**
  - ▷ allows syntactic integration (entity/relation framework)
  - ▷ powerful query language (SQL)
  - ▷ can cope efficiently with large datasets
  - ▷ allows query optimizations and thus intensive data mining
- **Each source has its own schema**
  - ▷ advantage : they can be updated individually
  - ▷ advantage : it is easy to add a new source of data
  - ▷ drawback : data are not reconciled...  
but redundancy helps in discovering complementarities and divergences between sources (DB cross-checking)
- **Semantic integration**
  - ▷ semantic *data* matching : link table
  - ▷ semantic *schema* matching : (multiple layer architecture)

# Microbiogenomics warehouse : architecture

- **The warehouse is built as a relational database (postgreSQL)**
  - ▷ allows syntactic integration (entity/relation framework)
  - ▷ powerful query language (SQL)
  - ▷ can cope efficiently with large datasets
  - ▷ allows query optimizations and thus intensive data mining
- **Each source has its own schema**
  - ▷ advantage : they can be updated individually
  - ▷ advantage : it is easy to add a new source of data
  - ▷ drawback : data are not reconciled...  
but redundancy helps in discovering complementarities and divergences between sources (DB cross-checking)
- **Semantic integration**
  - ▷ semantic *data* matching : link table
  - ▷ semantic *schema* matching : (multiple layer architecture)



# Microbiogenomics warehouse : data processing

- Primary data

- Parsing of ASCII flat files to create a relational model for each genomic collection
- Computation of the LinkTable between corresponding entities in the schemas

- Secondary data

- Search for homology relationships between proteins from all genomes with BLAST
- Extremely computer intensive (cross comparisons of millions proteins)
- Development of an incremental and parallel computation pipeline

- Tertiary data

- Development of specific programs to derive tertiary data from secondary data

# Semantic integration

## Querying the warehouse

# Querying the warehouse

## Querying/vizualizing individual source

- Most databases have their own Web querying/visualization interface
- Primary data, e.g., PAREO (KEGG metabolic pathways) ▶
- Secondary data, e.g., INSYGT/ORIGAMI (multiple-gene – multiple-genome navigator) ▶

## Querying the warehouse

- Microbiogenomics warehouse ~ 200 tables
- Sources with different schemas  $\implies$  complex relational schema
- Different entities (potentially in different sources) are involved in complex queries
- Formulating SQL queries on the warehouse is a difficult task for users

# Querying the warehouse

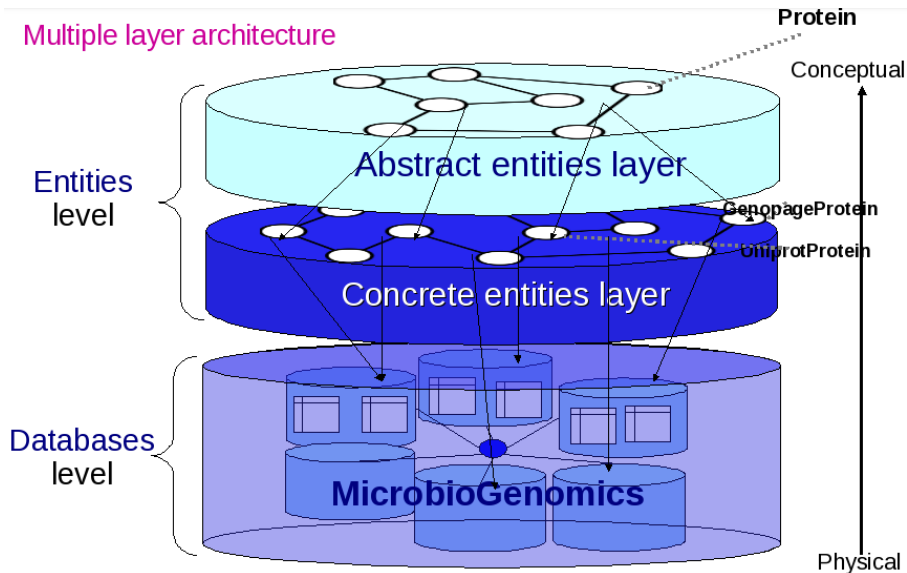
## Querying/vizualizing individual source

- Most databases have their own Web querying/visualization interface
- Primary data, e.g., PAREO (KEGG metabolic pathways) ▶
- Secondary data, e.g., INSYGT/ORIGAMI (multiple-gene – multiple-genome navigator) ▶

## Querying the warehouse

- Microbiogenomics warehouse ~ 200 tables
- Sources with different schemas  $\implies$  complex relational schema
- Different entities (potentially in different sources) are involved in complex queries
- **Formulating SQL queries on the warehouse is a difficult task for users**

# Semantic schema integration : conceptual framework

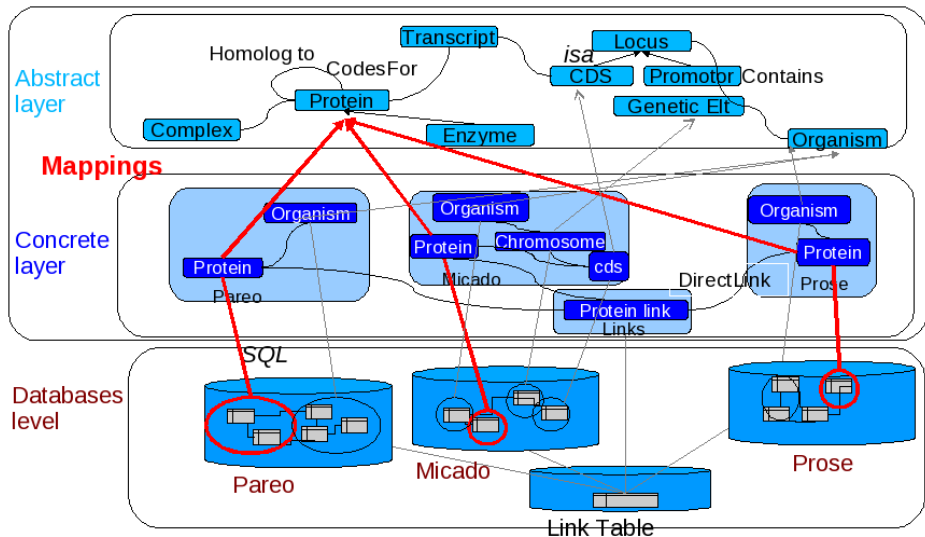


# Graphs of entities

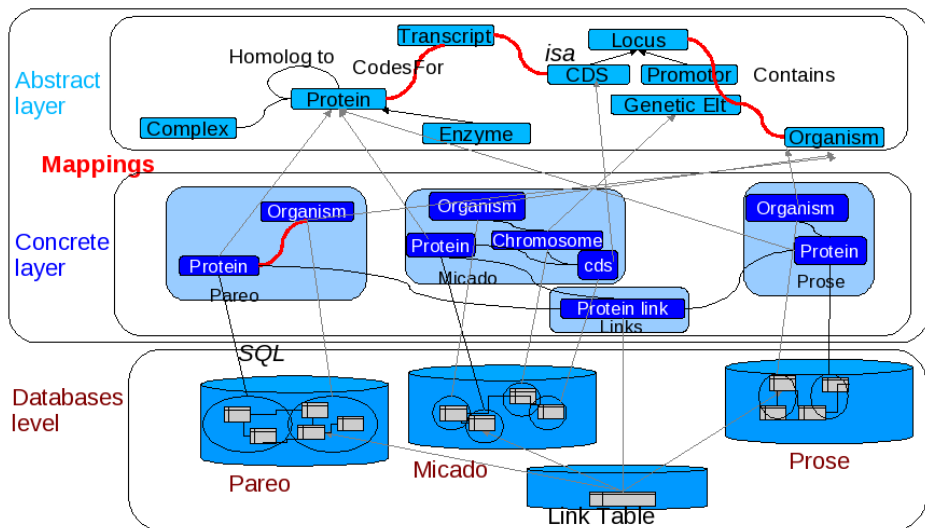
- Graph of **abstract** entities ▶
  - Two types of vertices :
    - biological entities, e.g., **Protein**, **MetabolicPathway**, **Organism**
    - properties, e.g., **Sequence**, **EC number**, **name**
  - Three types of edges :
    - isa, e.g., **Chromosome** ↔ **GeneticElement**
    - biological links, e.g., **CodeFor**, **Protein** ↔ **Transcript**
    - Has-For-Property, e.g., **Protein** ↔ **Sequence**
- Graph of **concrete** entities (views in the databases) ▶
  - Three types of vertices :
    - biological entities *in the sources*, e.g., **KEGGprotein**, **KEGGorganism**
    - properties *of the concrete entities*, e.g., **Sequence**
    - link entity (1 vertex : ProteinLink)
  - Three types of edges :
    - biological links, e.g., **ExistIn**, **KEGGprotein** ↔ **KEGGorganism**
    - Has-For-Property, e.g., **KEGGprotein** ↔ **Sequence**
    - Direct-Links, e.g., inter source connections between concrete entities
- Mapping between the level



# Conceptual framework : mapping



# Conceptual framework : mapping

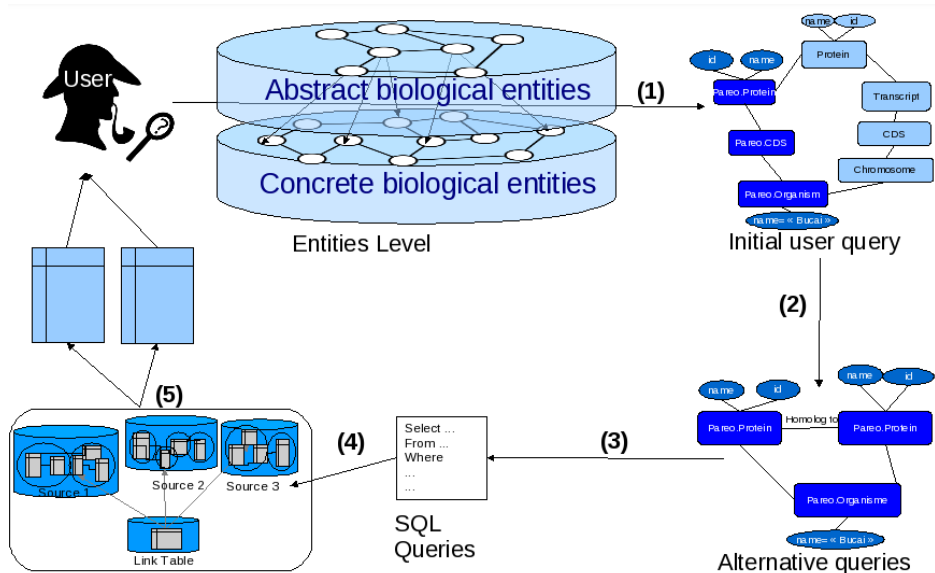


# Types of queries

One can pose queries involving :

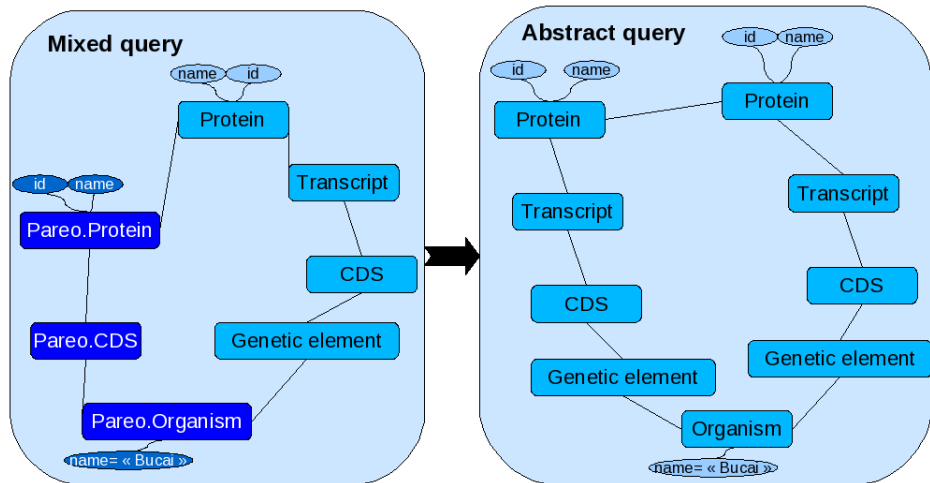
- abstract entities : **abstract** queries
  - ▷ Ex : What are the **enzymes** of the **organism** *Buchnera aphidicola* ?
- concrete entities : **concrete** queries
  - ▷ Ex : What are the **enzymes in KEGG** of the **organism** *B. aphidicola* **in KEGG**
- both : mixed queries
  - ▷ Ex : What are the **proteins** of the **organism** *B. aphidicola* **in KEGG**, and their **homologs (paralogs)** in the same **organism** ?

# Querying process



# Step 1 : abstraction

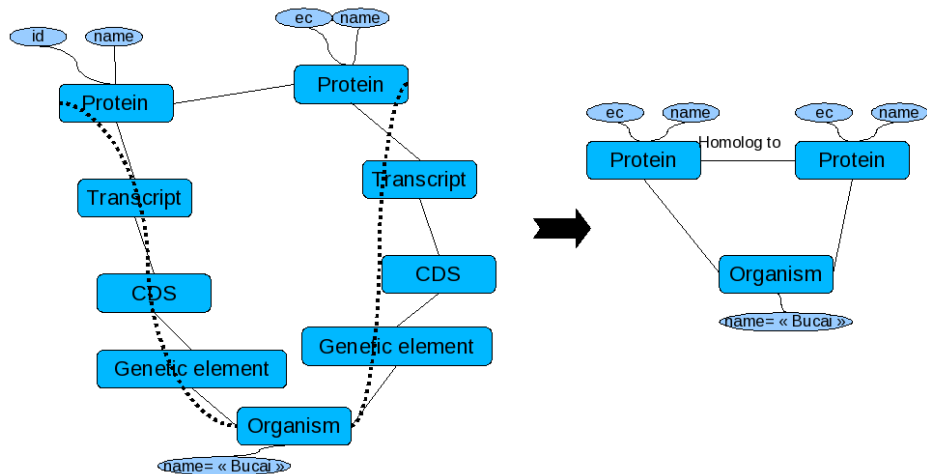
Find **proteins** of the **species** *Buchnera Aphidicola* that are in **PAREO (Kegg)** and have **homologs** in the same **species** :



## Step 2 : linking abstract entities of the query

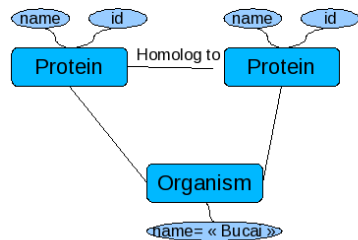
Input : abstract query

Output : high-level intermediate queries

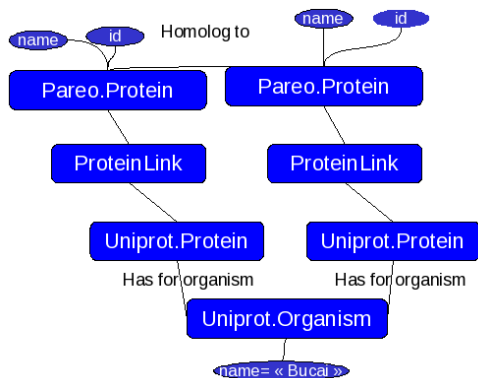


## Step 3 : generation of low-level queries

Input : high-level intermediate query

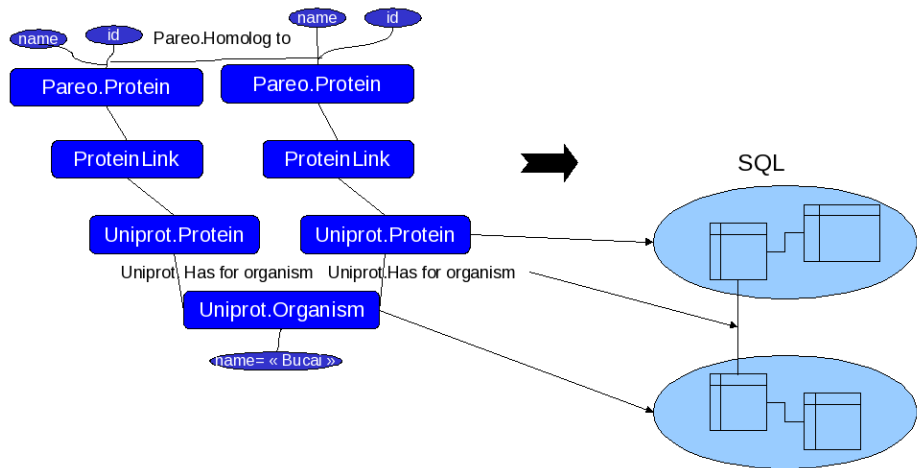


Output : alternatives queries



## Step 4 : automatic translation into SQL

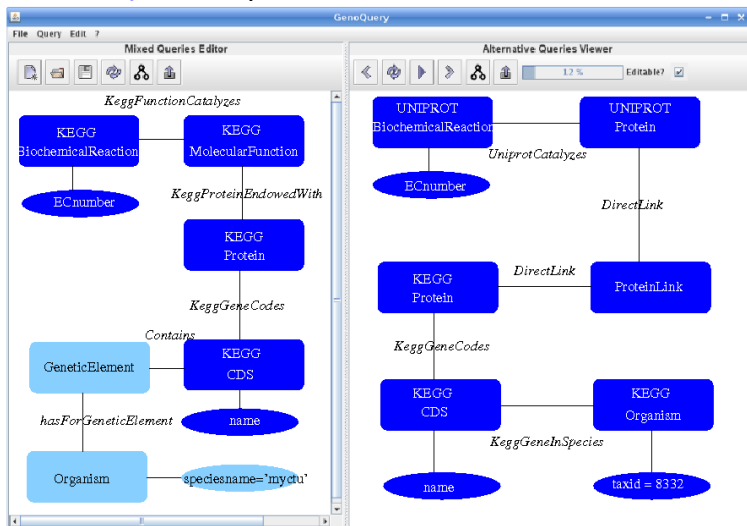
Evaluation on the databases by means of the mappings between the **concrete entity** layer and the **database** level





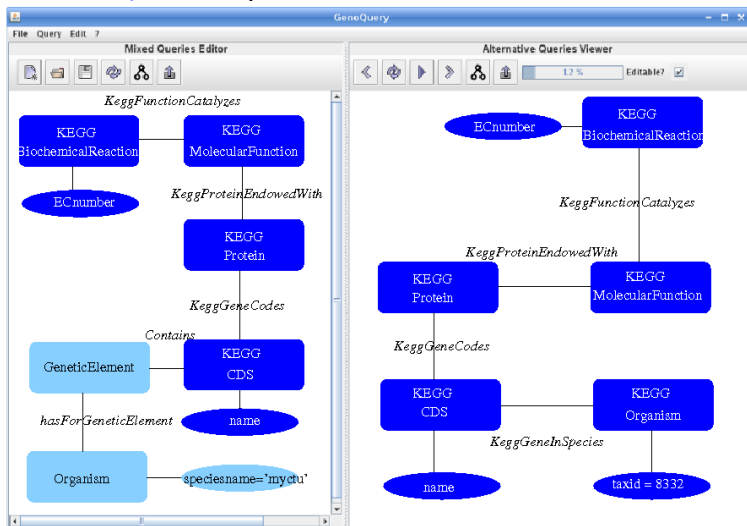
# GenoQuery : graphically querying the system

What are the **enzyme activities** (EC numbers) in **KEGG** catalyzed by **proteins** of the **species** *Mycobacterium tuberculosis*?



# GenoQuery : graphically querying the system

What are the **enzyme activities** (EC numbers) in **KEGG** catalyzed by **proteins** of the **species** *Mycobacterium tuberculosis*?



# Microbiogenomics warehouse : summary

- A *tightly integrated* and *materialized* solution for the integration of data sources required for microbial genome annotation, comparative genomics and gene evolutionary studies
- The warehouse is built as a relational data base made of independent schemas
  - Syntactic integration : entity/relation framework
  - Semantic integration :
    - ▷ data level : link table
    - ▷ schema level : a multi-layer architecture
      - ◇ A formalism based on graphs
      - ◇ A querying mechanism allowing the user to :
        - ↪ Graphically query the system (user-friendly interface)
        - ↪ Define mixed queries
        - ↪ Efficiently get answers from alternative queries

# Acknowledgements

## ▷ LRI

- C. Froidevaux
- F. Lemoine

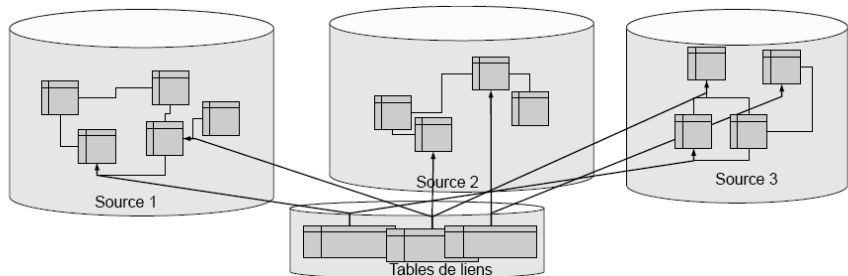
## ▷ IGM

- B. Labedan
- S. Descorps-Declere

## ▷ MIG

- V. Loux
- A. Gendrault
- T. Lacroix
- F. Papazian

# Link table



1 row = 1 protein = association of accession numbers (source IDs)

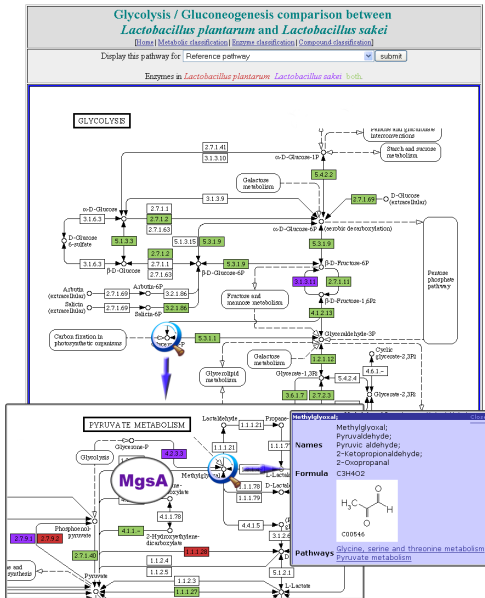
|              |             |             |             |     |
|--------------|-------------|-------------|-------------|-----|
| warehouse ID | Source 1 ID | Source 2 ID | Source 3 ID | ... |
|--------------|-------------|-------------|-------------|-----|

Procedure for computing the link table :

- sequence hash keys
- Blast comparisons
- database cross references

# PAREO (KEGG) interface

## PAREO



# INSYGT (synteny viewer)

welcome AgmialUser Menu

Search **Results Browsing** Genomic Organisation

**Gene Info**

Show Genomic Organisation

GENE INFO: glnA931  
[EC# (1.3), strand -]  
MATCHING INFO VS glnA30  
[Evalue : 0.0001  
For glnA931, start: 40, lenght: 46  
For glnA30, start: 32, lenght: 58]

SYNTENY INFO: group number2  
[group member : glnA9111 glnA981 g

ORGANISM INFO: Aeromonas hydrophila  
[ID# 10, Strain ATCC 7966 = NCIB 92

GENE INFO: glnA731  
[EC# (1.3), strand -]  
MATCHING INFO VS glnA30  
[Evalue : 0.0001

**Results Quick Navigation**

Options

**Bacillus subtilis**

<< Upstream Genes Set New Search Downstream Genes Set >>

5' glnA10 glnA20 glnA30 glnA40 glnA50 glnA60 glnA70 glnA80

glnA21 glnA41 glnA71 glnA81

Paging through Results : < previous 8 - 14 of 50 next >

**#10 Aeromonas hydrophila**

glnA910 glnA930 glnA960 glnA970 glnA980

glnA911 glnA931 glnA951 glnA961 glnA971 glnA981

**#11 Aeropyrum pernix**

glnA1010 glnA1030 glnA1040 glnA1070 glnA1080

glnA1011 glnA1031 glnA1041 glnA1061 glnA1071 glnA1081

Search **Results Browsing** Genomic Organisation

**Gene Info**

**Results Quick Navigation**

Bacillus subtilis [glnA10 glnA20 glnA30  
VERSUS  
Lactobacillus sakei [glnA3110 glnA380 c  
Bacillus subtilis [glnA10 glnA20 glnA30 glnA

Options

glnA30 glnA40 glnA50 glnA60 glnA70 glnA80 glnA90 glnA100 glnA110

glnA3100 glnA310 glnA390 glnA370 glnA330 glnA360 3'

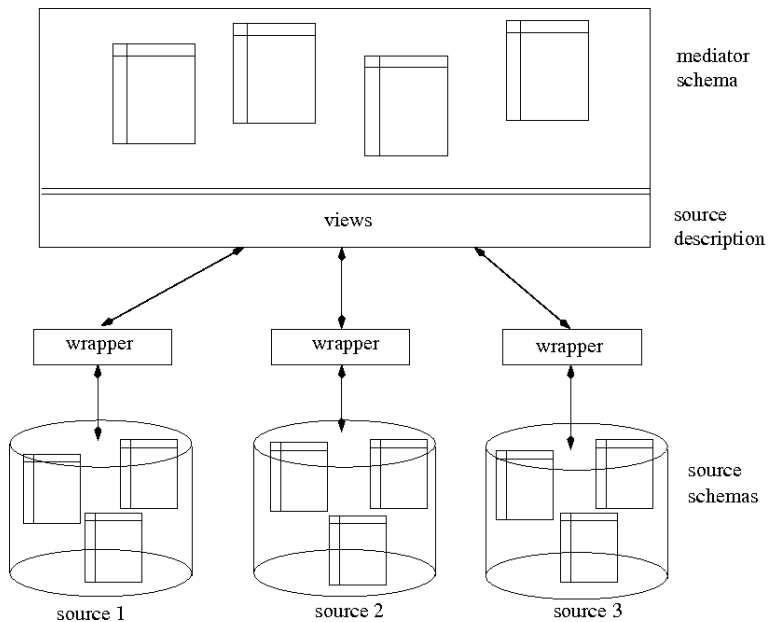




# Concrete entity graph



# Mediators



# Peer to peer

