Using BioCyc to reconstruct constraint-based models of metabolism : the Acinetobacter ADP1 case study

François Le Fèvre & Vincent Schachter

Genoscope, Evry, France

Genoscope's scientific focus

Genoscope :

- French national genome center
- Research institute

Themes :

- Genome annotation, genome evolution
  - Procaryote genome annotation
  - Paramecium annotation and genome evolution : reconstructing ancestral genomes
- Procaryote metabolism
  - Wet : efficient mutagenesis and phenotyping platform, biochemistry, adaptive evolution experiments
  - Dry : modeling and analysis of regulatory and metabolic networks

Some ongoing projects :

- ✓ « Metabolic Thesaurus » : elucidating the metabolism of a model bacterium, Acinetobacter ADP1
- ✓ Adaptive evolution experiments on bacteria
- ✓ « Project Biodiversity - Cloaca maxima » : analysis of a large bacterial metagenome (samples from water retreatment facility) -> bacterial diversity and biodegradation

Outline

- Context : the Metabolic Thesaurus Project
- Reconstructing a global model of Acinetobacter ADP1 metabolism
  - Ab-initio reconstruction of a global metabolic model
  - Initial model refinement using phenotype experiments and predictions
  - Use of BioCyc as a “middleware”
- Ongoing development of OOcyc
  - an API (+ RDB + O/R mapping) aimed at querying/ manipulating BioCyc information in an object oriented manner
- Conclusions & future work

Outline

- Context : the Metabolic Thesaurus Project
- Reconstructing a global model of Acinetobacter ADP1 metabolism
  - Ab-initio reconstruction of a global metabolic model
  - Initial model refinement using phenotype experiments and predictions
  - Use of BioCyc as a “middleware”
- Ongoing development of OOcyc
  - an API (+ RDB + O/R mapping) aimed at querying/ manipulating BioCyc information in an object oriented manner
- Conclusions & future work

The Metabolic Thesaurus Project

- **Bacterial metabolism is very diverse and not well known outside of E.coli**
  - What can we learn of the metabolism of a new model bacterium ?
  - Can metabolic modeling help ?

- **Experimental context :**
  - Reliable genome annotation
  - Medium/high throughput mutant construction capabi
  - Medium/high throughput phenotyping capab
  - Additional experiments on an ad-h
    - Enzyme kinetics
    - Gene expressio
    - Metaboli

WORK IN PROGRESS

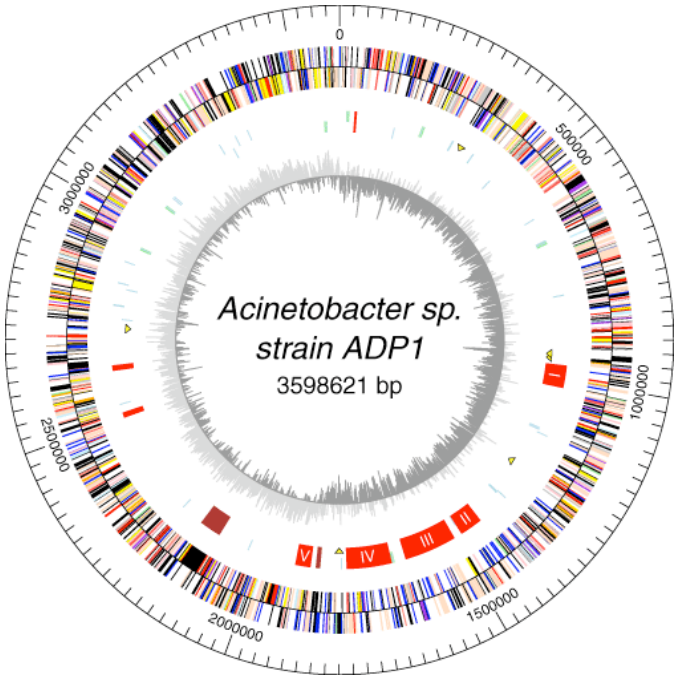
- **Model**
  - Rec truction of global metabolic network model
    - Initial model reconstruction
    - Correct/complete model iteratively using phenotype experimental results
    - Reach best fit between phenotype experiments and phenotype predictions
  - Improve modeling framework to account for regulatory effects
  - Assess information content and structure of large-scale phenotype data
  - Help with experiment design

Choosing a model organism : Acinetobacter sp. ADP1

- Acinetobacter sp. ADP1
  - γ-proteobacteria, Pseudomonales group
  - Gram negative
  - Soil bacterium
  - Non flagellated

- A good model for the study of procaryote metabolism
  - Nutritionally versatile
  - Strictly aerobic
  - Fast growing
  - Highly competent (natural transformation)
  - Non-pathogenic
  - Fairly close to E.coli, closer to Pseudomonas putida
  - Recent evidence of xenobiotic degradation capabilities

Genome Annotation



Expert annotation of *Acinetobacter* sp. ADP1 : 3201 CDS **Barbe et al., 2004**

Known genes :	**1150 genes**
Putative genes :	**907 genes**
Conserved Hypothetical protein :	**686 genes (21%)**
Hypothetical proteins:	**458 genes (14%)**

36% of CDS have unknown functions

Collection of gene replacement mutants of *Acinetobacter* sp. ADP1:

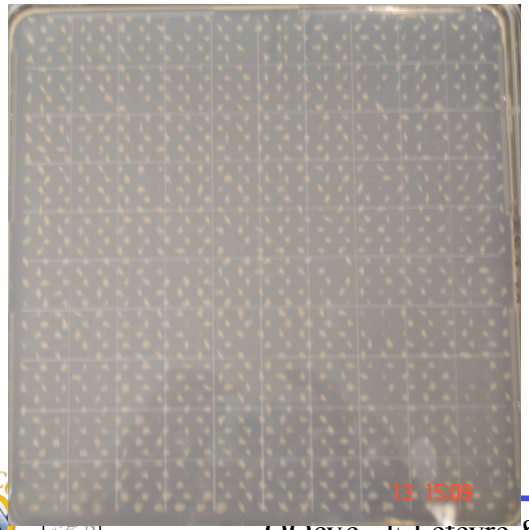
Systematic gene knockout project using a gene replacement procedure consisting of an excision of the target gene via homologous recombination and insertion of a kanamycin resistance marker (Metzgar et al., 2004)

Status of the collection:

Known genes :	**684 KO**	**(59% of all known genes)**	**2150 KO**
Putative genes :	**736 KO**	**(81% of all putative genes)**	
Conserved hypothetical protein :	**482 KO**	**(70% of all CHPs)**	
Hypothetical proteins:	**247 KO**	**(54% of all HPs)**	

~300 KO could not be obtained on succinate.
Almost all are known to be essential in *E. coli*

Systematic phenotyping of the collection on carbon sources



Currently :~25000 phenotype data points
18 carbon sources + 2 nitrogen source on 1350 KO

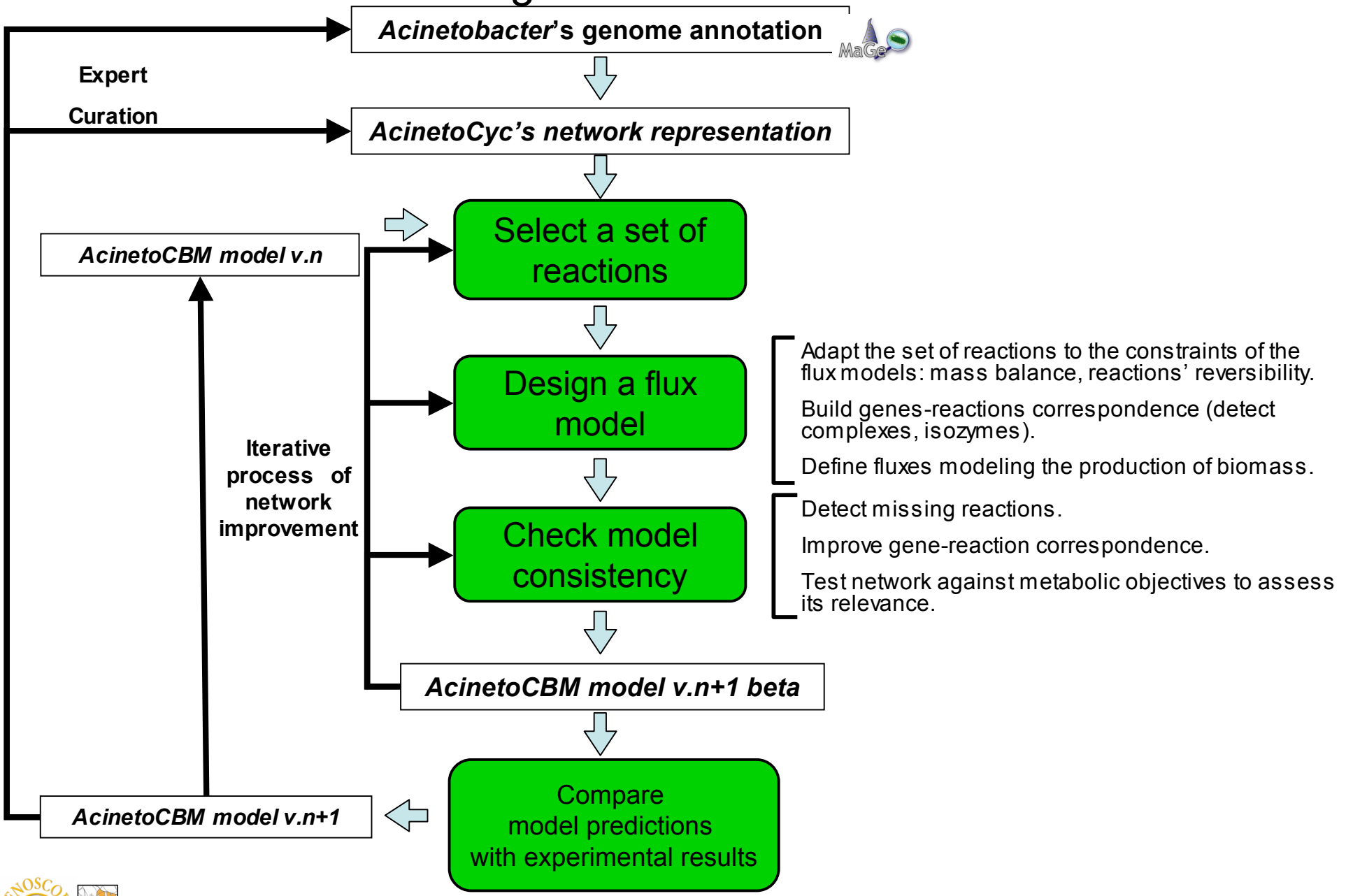
Under development :

- **high-throughput phenotyping on the collection for ~50 C/N/S sources**
- **1200 KO/plate**

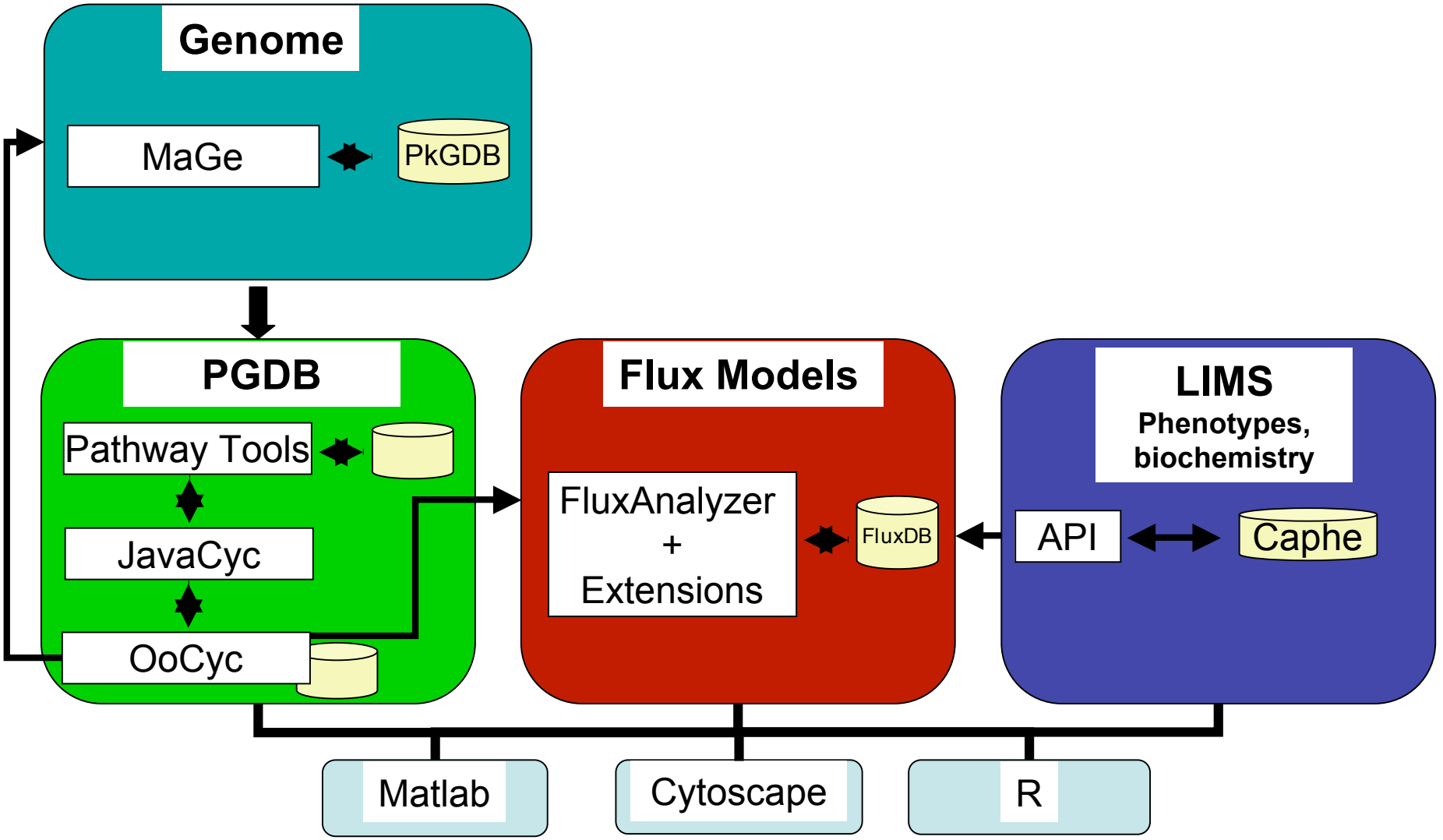
The constraint-based modeling framework (aka global stoichiometric models)

- Variables of interest : metabolite fluxes
  - State of the system : distribution of fluxes through all reactions
- Steady-state assumption
- Key ideas :
  - Reason on sets of metabolic states rather than on unique « solution » states
    - Allows for incompleteness or inaccuracies in both the reaction network structure and the experimental constants
    - The feasible solution space can be progressively reduced using constraints (thermodynamic, input/output, regulatory…)
  - Allows prediction of optimal (or good) flux distributions given
    - an optimization principle
    - a defined metabolic objective (or several objectives)
  - Characterize structural features of flux distributions resulting from constraints
    - Evolutionary studies
    - Correlation with regulatory effects

Iterative building of Acinetobacter's flux model



Reconstruction infrastructure



Acinetocyc curation

Genes	3428
Pathways	234
Enzymatic Reactions	1198
Transport Reactions	3
Polypeptides	3377

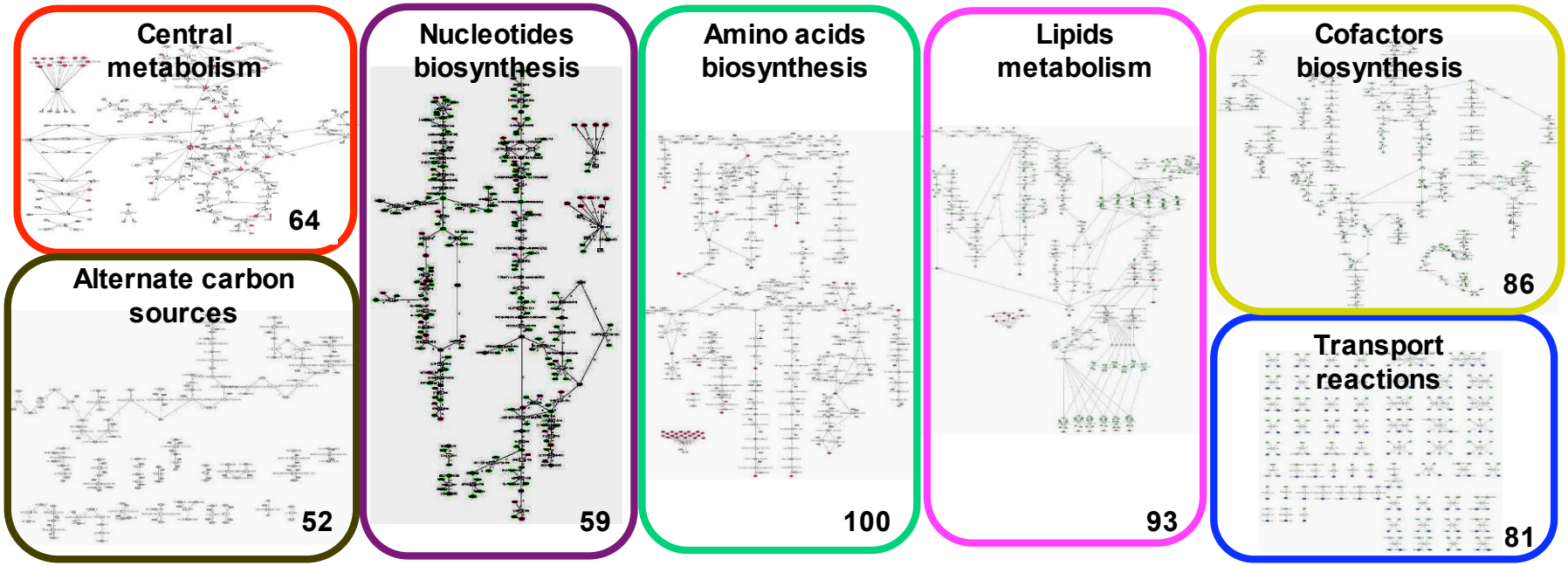
Protein Complexes	8+276
Enzymes	825
Transporters	3 !
Compounds	864
Transcription Units	2274
tRNAs	76

Pathways added in the curation process

- Entner-Doudoroff,
- Entner-Doudoroff Neoglucognesis Pentose Phosphate Superpathway,
- Degradation of glucarate and galactarate,
- Ferulate catabolism,
- Caffeate catabolism,
- Superpathway of hydroxycinnamates catabolism,
- …,

with corresponding new compounds and reactions.

Reconstruction of Acinetobacter's metabolic network

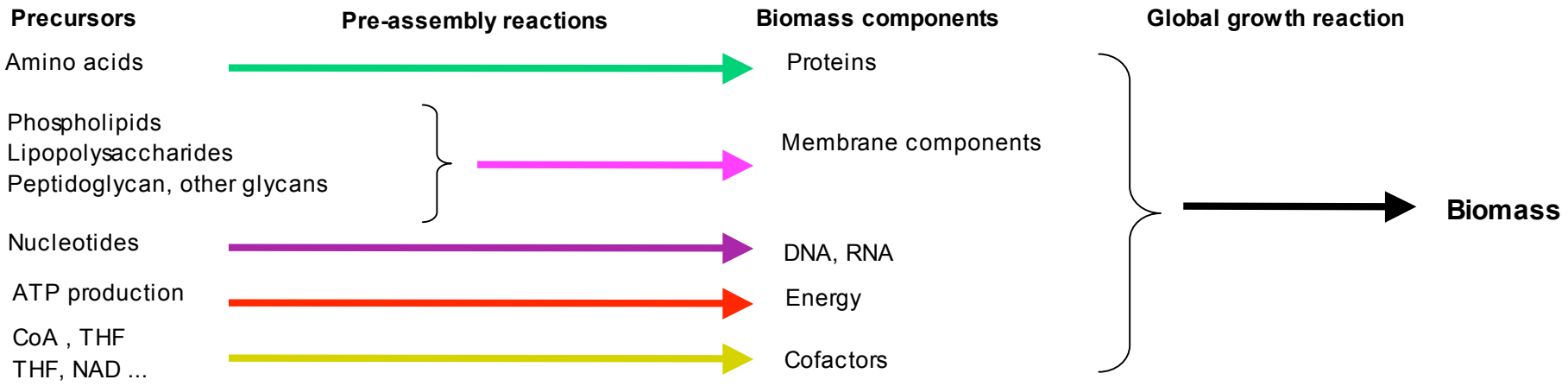


535 reactions using **498** metabolites linked with **480** genes dispatched in **7** functional sub-networks.

297 enzymatic complexes have been inferred by naïve comparison with *E. coli.* (Draft version: expert curation to be performed)

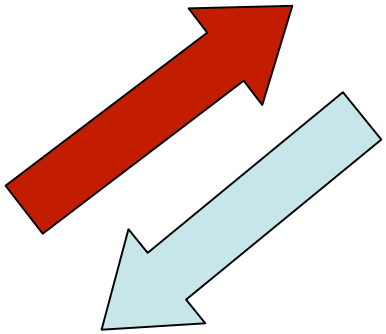
Reconstruction of Acinetobacter's metabolic network

– Simulate growth by a reaction using the precursor metabolites of biomass



– Optimize the flux distribution to **maximize** this growth flux

Constraint-based models are not deduced trivially from PGDBs (& vice-versa !)…



AcinetoCyc

- Reactions with no corresponding gene
- Unbalanced reactions
- **Non-metabolic reactions**
- Reactions with undefined compound(s)
- Reactions with generic compound
- Reactions with undefined reversibility

Well defined reactions

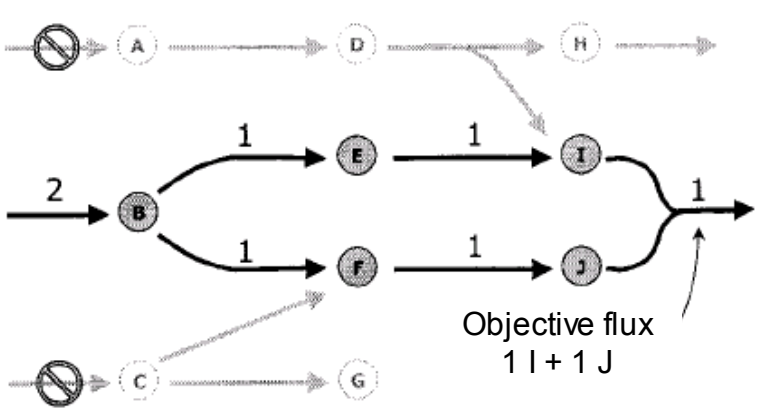
Flux model

- Transport reactions
- **Biomass reaction(s)**
- **Precursor assembly reactions**
- Polymerization reactions

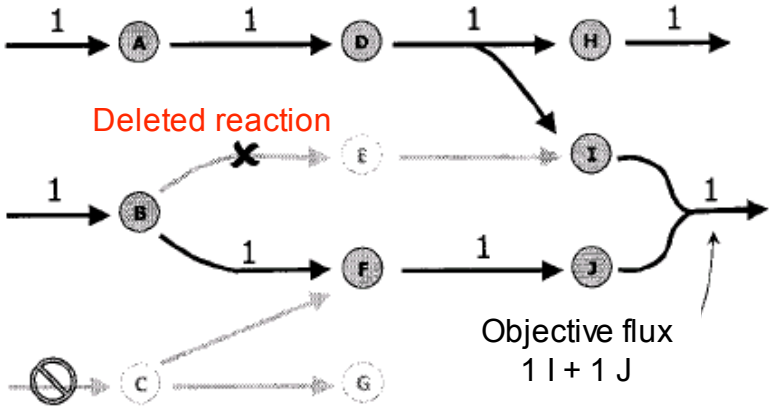
Prediction of mutant phenotypes

- Principle: for each mutant phenotype “experiment”,
  - determine which reaction to block (simulate the knock-out),
  - determine exchange fluxes to open (simulate the growth medium),
  - optimize biomass flux.

Wild type network



Mutant network



- Growth on B alone is possible.

- Growth on B alone is impossible
- Growth on A and B is possible.

Prediction of mutant phenotypes

- Discretization of predicted biomass flux values :
  - » no growth (-),
  - » slow growth relative to the wild type (+),
  - » normal growth (++).

- Current dataset :
  - 480 *in silico* KO (metabolic genes)
  - 1300 experimental KO + 195 genes suspected to be essential

  ➔ overlap of **186** genes.

Analysis of local inconsistencies

34 inconsistencies with first set of experiments x predictions

Gene	Initial enzymatic activity	akGl	Acet.	Gluc.	Lact.	Pyru.	Succ.
0476	Asparaginase	0/1	0/1	0/1	0/1	0/1	0/1
2879	Succinate dehydrogenase	0/1	0/1	1/1	1/1	1/1	0/1
0272	Glutamate--tRNA ligase	1/1	1/1	1/0	1/1	1/0	1/0
3371	Glutamate--tRNA ligase	1/0	1/0	1/0	1/0	1/0	1/0
2875	Oxoglutarate dehydrogenase	1/0	1/0	1/0	1/0	1/0	1/1
2876	Oxoglutarate dehydrogenase	1/0	1/0	1/0	1/0	1/0	1/1
0546	G3P dehydrogenase	1/1	1/1	1/0	1/1	1/1	1/1
2893	P-Pantetheine adenylyltransferase	1/0	1/0	1/0	1/0	1/0	1/0

Possible explanations

- Missing isozyme, later confirmed
- Detection of experimental error: gene was not deleted
- Hypothesis: pair of isozymes, constitutive versus regulated expression
- Evidence for the existence of a complex rather than isozymes
- Detection of experimental error: gene in Entner-Doudoroff operon
- CoA had to be added to biomass production, forcing completion of co-factor synthesis pathways

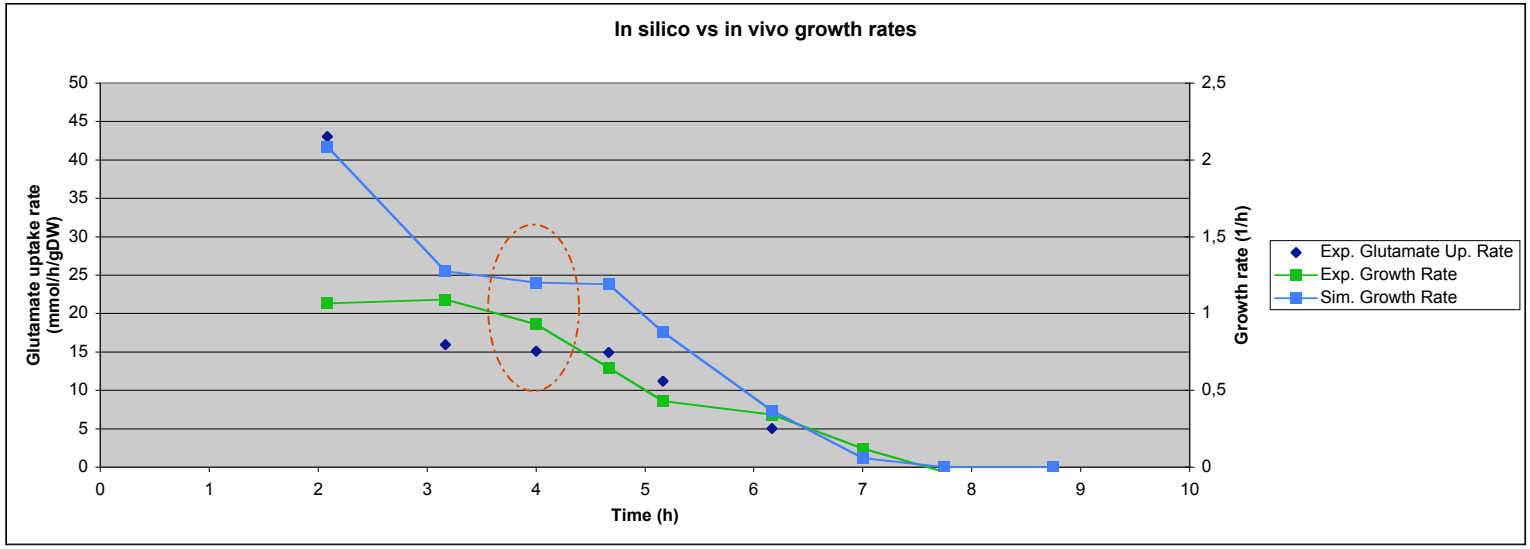
IS: *in silico* / IV: *in vitro*.

0 : 0 → no growth 1 : WT value → normal growth

0/1 False negative

1/0 False positive 1/1 True positive

Predictive quantitative growth : preliminary results

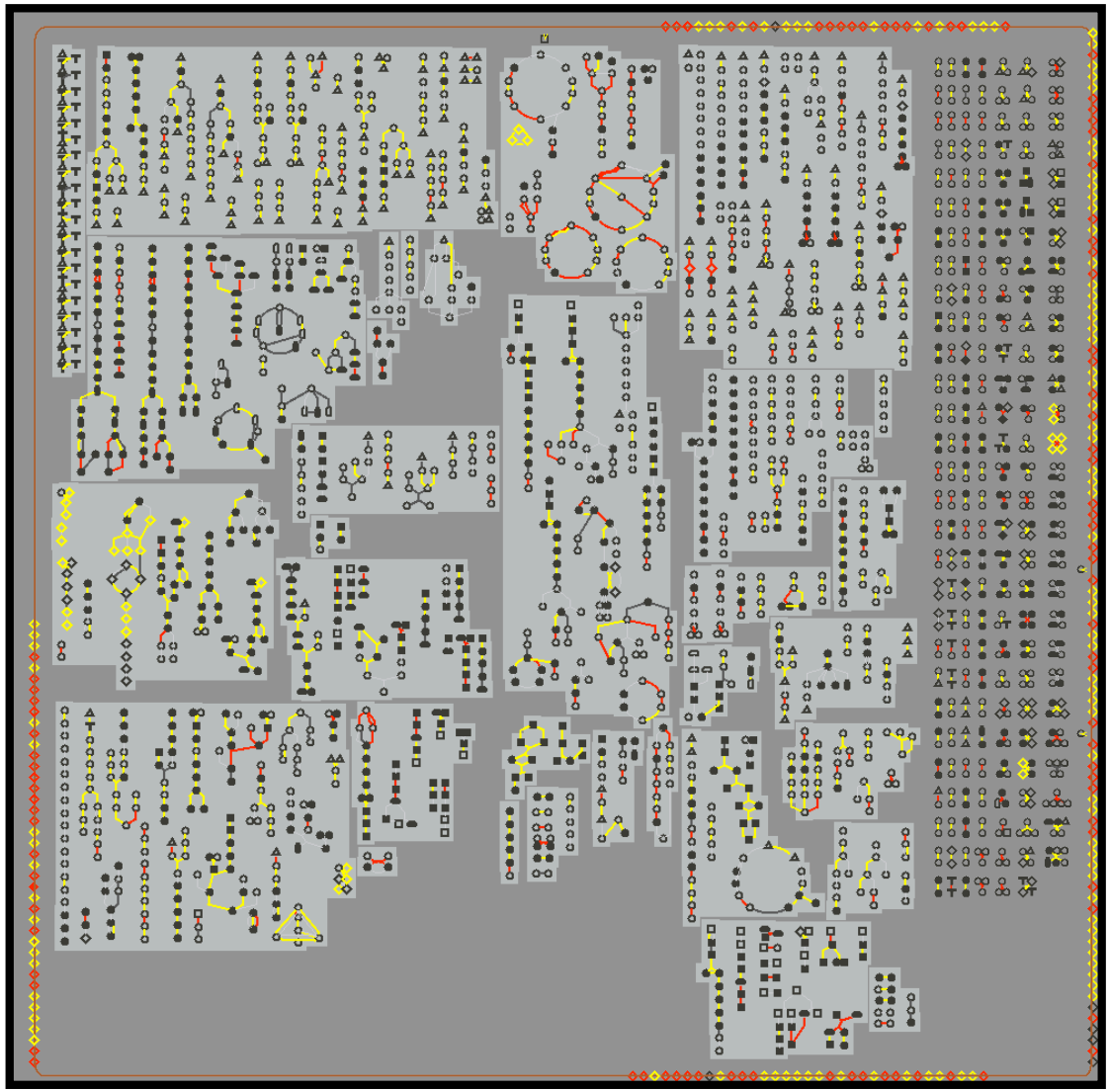


Experiment :

- aerobic growth of wt Acinetobacter on glutamate + nh4 + minimal media
- carbon source consumption and biomass production

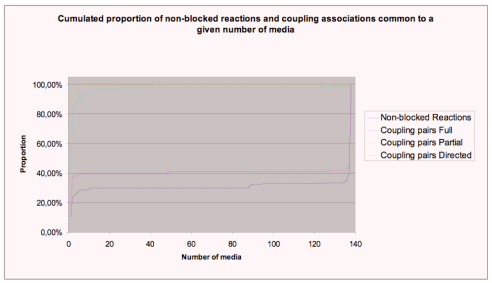
- Preliminary but encouraging results
- More experimental data coming…
- Possible improvements :
  - Refine energy consumption-related pathways
  - Refine biomass composition

Visualizing the progress of growth phenotype assays



Ongoing related work

- Development of methods to learn metabolic models from experimental data
  - Use phenotype profiles systematically to refine models
  - Extend modeling framework to exploit information on metabolite concentration
- Metabolism & regulation ?
- Assessment of flux coupling graph variability with environmental change
  - Most coupling relationships are either very specific or very generic
  - Some coupling relationships exhibit more suprising patterns of occurrence



- Identify new enzymatic functions from a wastewater retreatment metagenome
  - species inventory , enzymatic activities inventory, gene inventory
    - Estimate : ~1000 species, mostly uncultured
  - assessment (some) metabolic processes in wastewater retreatment
  - those enzymatic functions that can't be found by straigthforward homology…

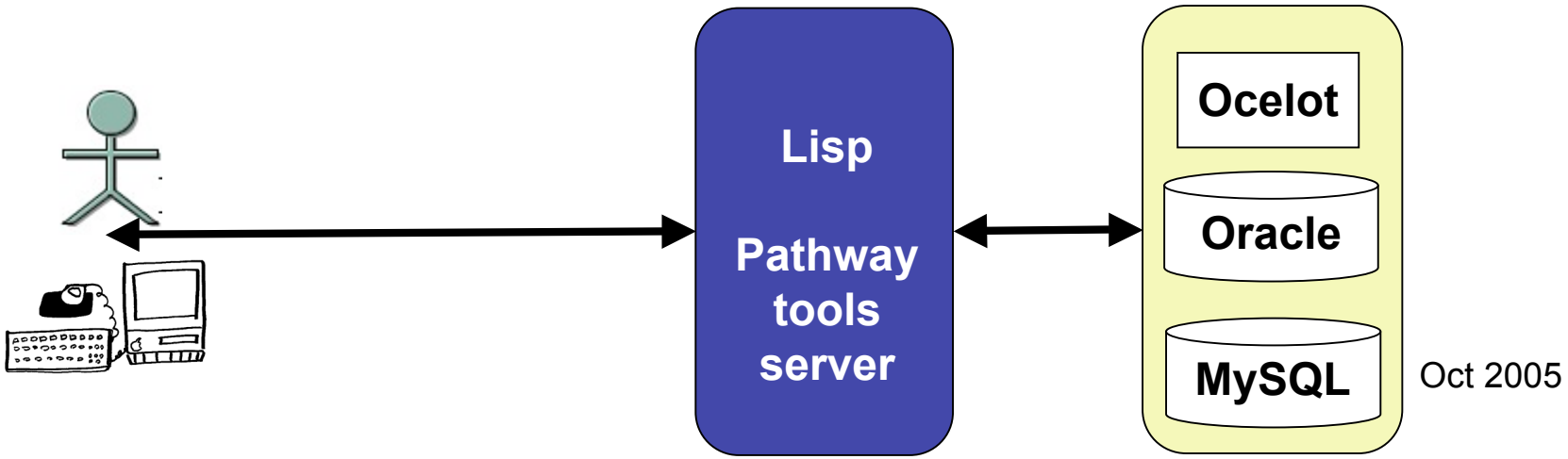
Outline

- Context : the Metabolic Thesaurus Project
- Reconstructing a global model of Acinetobacter ADP1 metabolism
  - Ab-initio reconstruction of a global metabolic model
  - Initial model refinement using phenotype experiments and predictions
  - Use of BioCyc as a “middleware”
- Ongoing development of OOcyc
  - an API (+ RDB + O/R mapping) aimed at querying/ manipulating BioCyc information in an object oriented manner
- Conclusions & future work

Existing means to query/manipulate BioCyc information
(as far as we know !)

- Graphical User Interface
  - Web-based
  - PathwayTools Interface
- Through the Pathway tools server
  - Lisp API
  - PerlCyc
  - JavaCyc
- Exports
  - text files, attribute-value or tab-delimited
  - SBML format
  - BioPAX format
- Biowarehouse
  - SQL queries

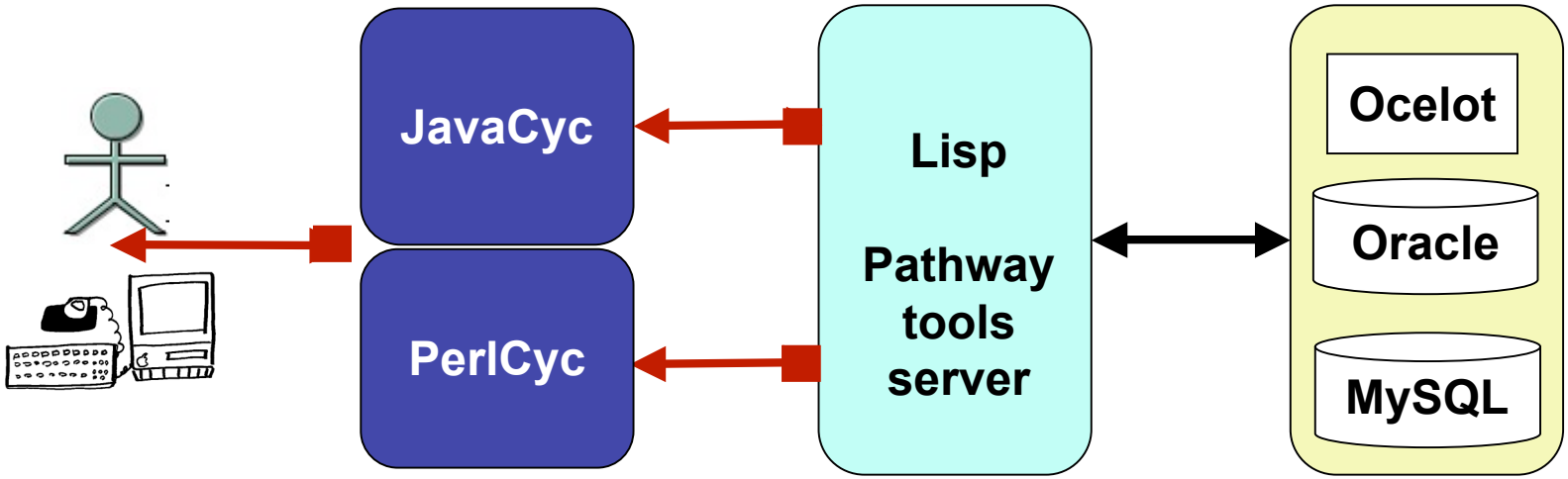
Accessing BioCyc information : the Lisp API



- Lisp is the native language of Pathway-tools,
- Multi-user editing (RDBMS versions)
- Access to a transaction log of all PGDB edits

- Need to know Lisp
- Low-level mySQL database schema → SQL queries are…tedious

Accessing BioCyc information : the JavaCyc and PerlCyc APIs



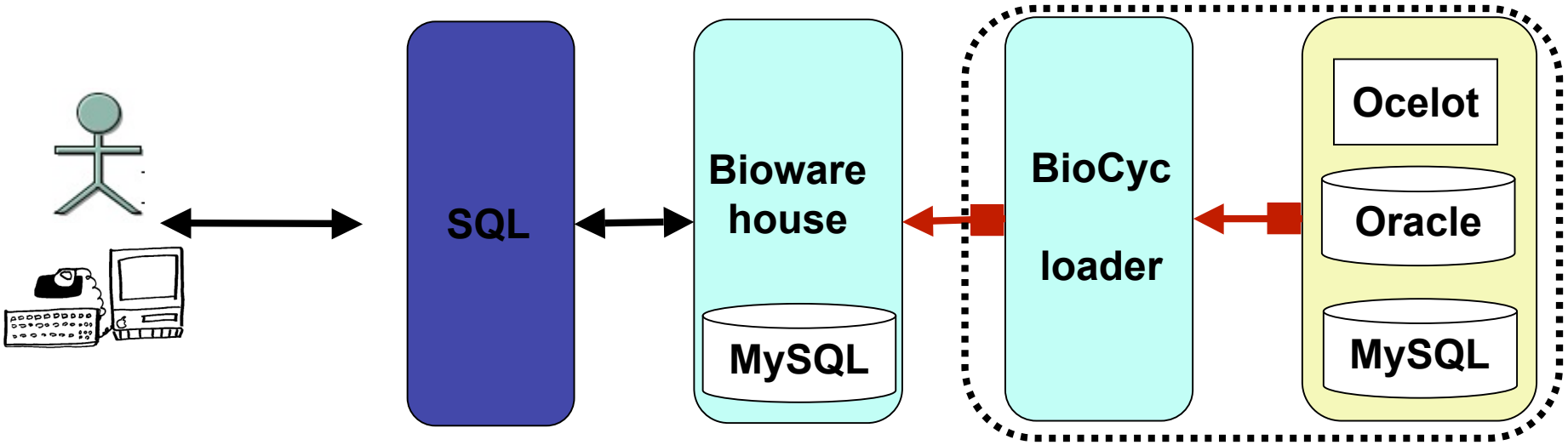
+

- Solutions for Lisp function access through java or perl.

-

- No representation of GFP entities in Perl or Java
- « Read-only » (JavaCyc)
- Need to run a special socket server
- Only one such connection can be opened
- OS dependant : Unix only

Accessing BioCyc information : the Biowarehouse API



+
- relational schema closer to biological concepts
- a solution to query BioCyc info integrated with external data
- simple installation
- fast data upload

-
- « read-only »
- loss of information wrt BioCyc model (transcription unit, citation, compound-pathway classification, functional category …)
- querying is a bit tedious

Motivations for OOcyc development

Within the context of Acinetocyc curation and Acinetobacter's CBM reconstruction :

- Overcome some limitations of BioCyc + Pathway tools
  - OS-dependency (JavaCyc : Unix)
  - no concurrent access (until recent MySQL version)
  - memory management (load entire PGDB in memory)

  before release 9.5…
- Manipulate objects close to natural biological representations via (non-LISP) queries
- Avoid loss of information with respect to Biocyc's model
- Add import / export features
- Benefit from existing Java tools / packages

What is OoCyc ?

- OoCyc is an Object Oriented API aimed at accessing and manipulating BioCyc information
- OoCyc stores information in a relational database in (your favorite) RDBMS
- OoCyc maps :
  - the FRS model onto an OO representation (& vice-versa)
  - frame instances onto objects (& vice-versa)
- OoCyc imports / exports Biocyc information (frame instances) from/to XML files (serialized objects)

From Frames …

… to Java classes

BioCyc

Frame RXN-1

Slot Parents

Value GeneralizedReactions

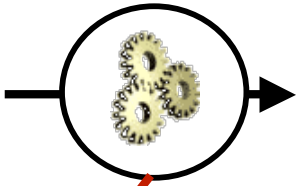
Slot Left

Value Glucose

Annotation1 Compartment

Annotation1.Value Cytoplasm

Slot …



OoCyc

GeneralizedReactions

Id=RXN-1

…

Left

Value=Glucose

Compartment=Cytoplasm

…

Chemicals

Id=glucose

…

Mapping rules

Top level parent → **Class**

Inheritance path → association classes

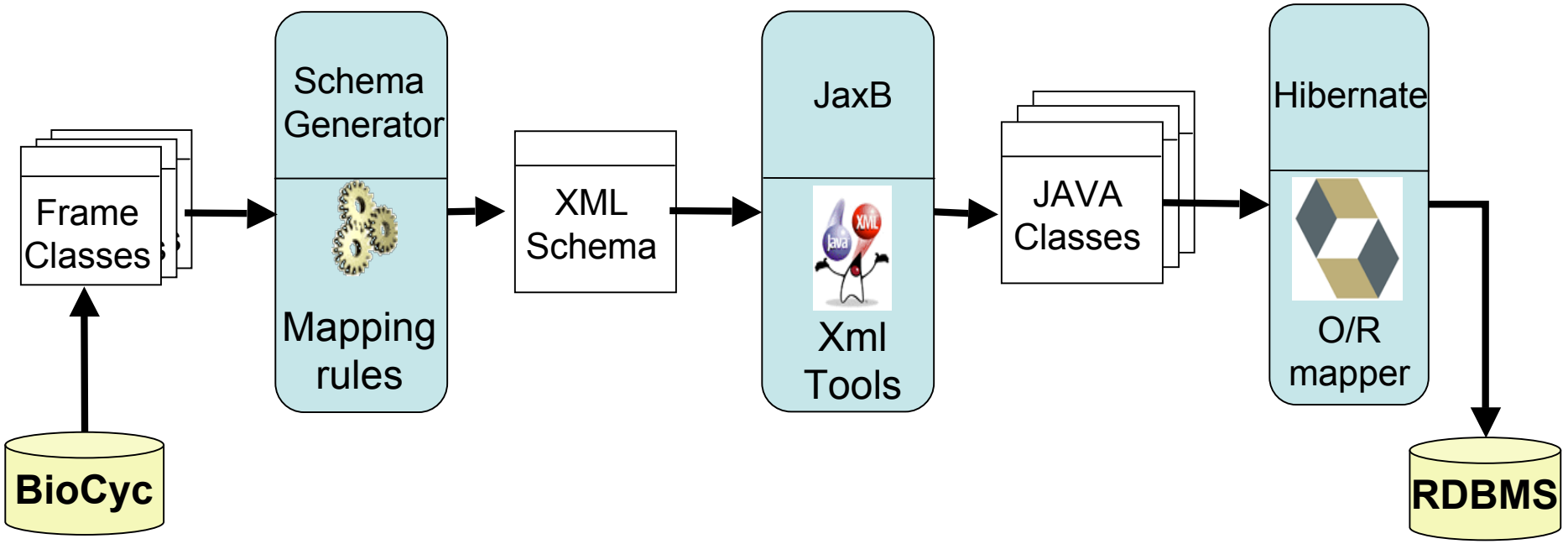
Slot → **Association class**

Slot Value → Attribute value

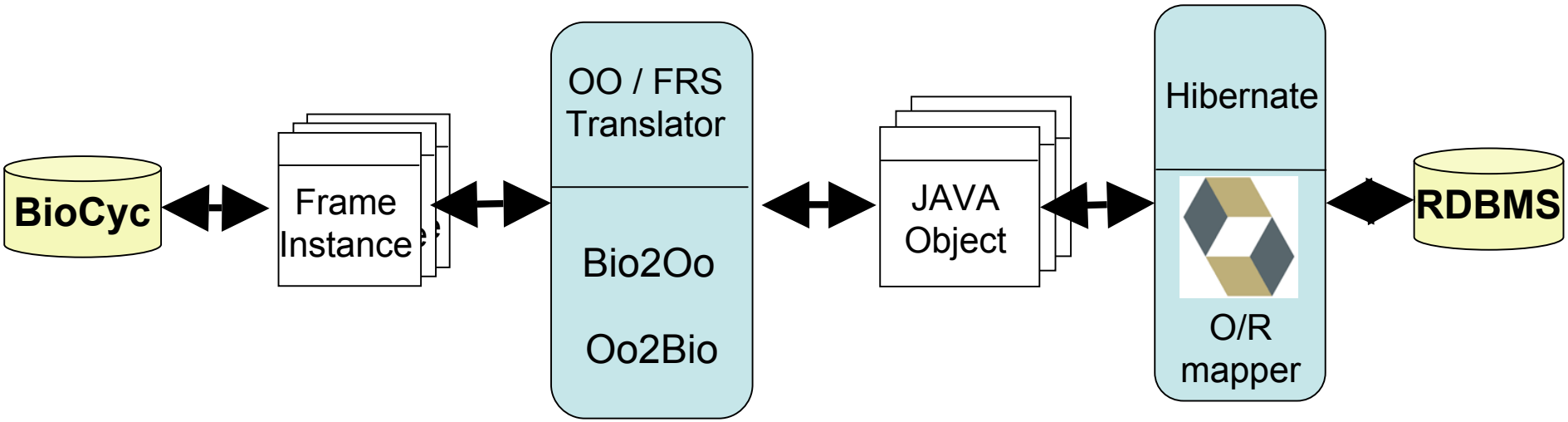
Annotation → Attribute name

Annotation.value →Attribute value

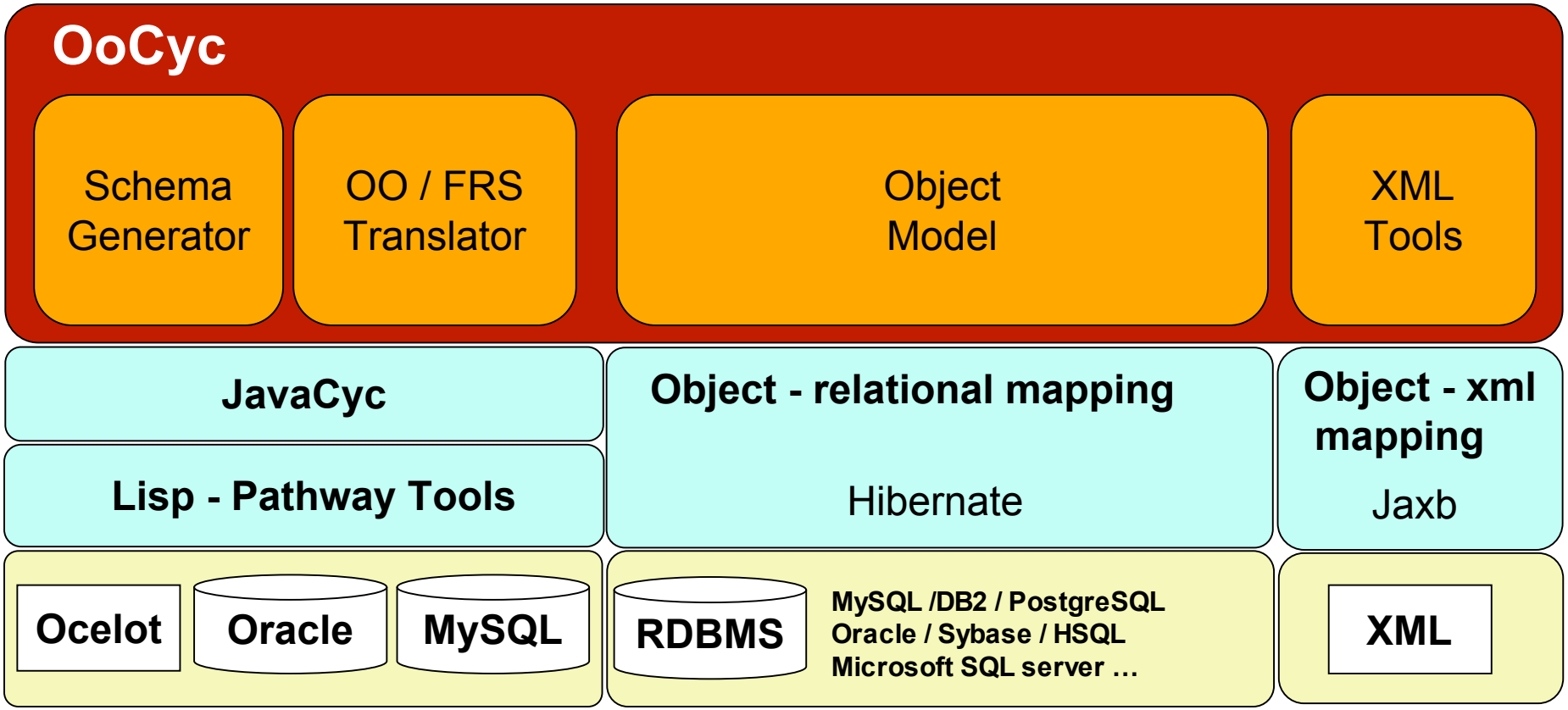
OoCyc's pipeline - initial step: code and database generation



OoCyc's pipeline - daily step: data synchronization

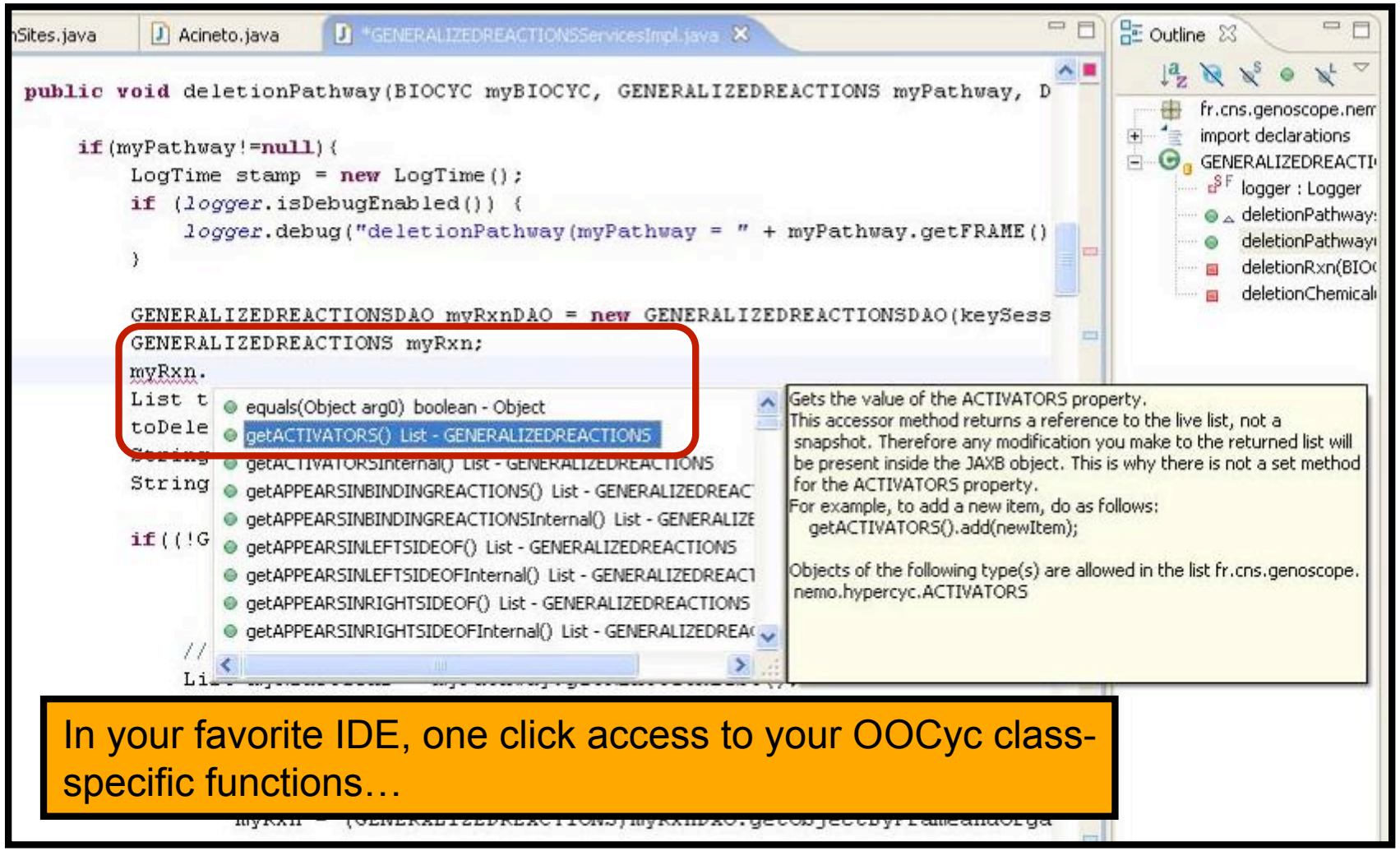


OoCyc's software architecture



Data, Files | Software, tools | Genoscope

Beneficits of using OoCyc : Java API matching biological concept



Beneficits of using OoCyc
Query in OQL (Object Query Language)

Lisp : Find all enzymes for which ATP is an inhibitor

```
1    (defun atp-inhibits ()
      ;; We check every instance of the class
2     (loop for x in (get-class-all-instances |Enzymatic-Reactions|)
      ;; Test for whether the INHIBITORS-ALL slot contains the
                                                  compound frame ATP
3      when (member-slot-value-p x INHIBITORS-ALL ATP)
       ;; Whenever the test is positive, we collect the value of the
       slot ENZYME The collected values are returned as a list, once
                                                  the loop terminates.
4       collect (get-slot-value x 'ENZYME))
5    )
     ;; invoking the query:
6    (select-organism :org-id 'ECOLI)
7    (atp-inhibits)
```

Beneficits of using OoCyc
Query in OQL (Object Query Language)

PerlCyc : Find all enzymes for which ATP is an inhibitor

```
1   use perlcyc;
2   my $cyc = perlcyc -> new("ECOLI");
3   my @enzrxns = $cyc -> get_class_all_instances("|Enzymatic-
                                                  Reactions|");
    ## We check every instance of the class
4   foreach my $er (@enzrxns) {
    ## We test for whether the INHIBITORS-ALL slot contains the
                                          compound frame ATP
5       my $bool = $cyc -> member_slot_value_p($er,"Inhibitors-
                                                All", "Atp");
6       if ($bool) {
          ## Whenever the test is ≥0, we collect the value of
          the slot ENZYME. The results are printed in theterminal
7          my $enz = $cyc -> get_slot_value($er, "Enzyme");
8          print STDOUT "$enz\n";
9       }
10  }
```

Beneficits of using OoCyc
Query in OQL (Object Query Language)

JavaCyc : Find all enzymes for which ATP is an inhibitor

```
1  import java.util.*;
2  public class JavacycSample {
3  public static void main(String[] args) {
4   Javacyc cyc = new Javacyc("ECOLI");
5   ArrayList enzrxns = cyc.getClassAllInstances("|Enzymatic-
                                                  Reactions|");
6     for (int i = 0; i < enzrxns.size(); i++) {
7       String er = (String)enzrxns.get(i);
8       boolean bool = cyc.memberSlotValueP(er, "Inhibitors-
                                    All","Atp");
9       if (bool) {
10        String enz = cyc.getSlotValue(er, "Enzyme");
9         System.out.println(enz);
11      }
12     }
13   }
14  }
```

Beneficits of using OoCyc
Query in OQL (Object Query Language)

Biowarehouse : Find all enzymes for which ATP is an inhibitor

```
1   SELECT DISTINCT DBID.xid
2   FROM DBID, Protein, EnzymaticReaction, Chemical, DataSet,
                              EnzReactionInhibitorActivator
3   WHERE DataSet.name=EcoCyc
4   AND DataSet.wid=EnzymaticReaction.datasetwid
5   AND EnzymaticReaction.proteinwid = Protein.wid
6   AND EnzymaticReaction.wid =
              EnzReactionInhibitorActivator.enzymaticreactionwid
7   AND EnzReactionInhibitorActivator.compoundwid=Chemical.wid
8   AND EnzReactionInhibitorActivator.inhibitoractivate=I
9   AND Chemical.name=ATP
10  AND DBID.otherwid = Protein.wid
```

Beneficits of using OoCyc
Query in OQL (Object Query Language)

- OQL is a powerful and easy-to-use SQL-like query language
  - object-oriented, understanding notions like inheritance, polymorphism and association.

OoCyc : Find all enzymes for which ATP is an inhibitor

```
1 SELECT reaction FROM ENZYMATICREACTIONS reaction
2  LEFT JOIN reaction.INHIBITORSALLInternal inhibitors
3  WHERE reaction.ORGANISM.VALUE LIKE :organism
       AND
4  "ATP" = inhibitors.VALUE
```

Summary of representation & query solutions

	BioCyc		Biowarehouse	OoCyc
Representation System	FRS		Relational model	**Object** model
Storage	Ocelot	mySQL oracle	mySQL	mySQL , Oracle Sybase, PostgreSQL …
Query	Lisp	SQL	SQL	**OQL** SQL Java
API for :	Lisp, Perl, Java		-	**Java**
BioCyc Compatibilty	-		Partial	**Full** **(but not perfect…)**
Import / Export to BioCyc	-		Yes / No	**Yes / Yes**
Independent from Pathway tools	No		Yes	**Yes**

What did we use OoCyc for ?

- **Extract data from BioCyc**
  - to integrate them with other data types
    - clusters of transcription units, pathways, complexes, Regulon DB..
  - to automate more steps of constraint-based building :
    - group together Pathways for map design in Flux Analyzer,
    - assemble elementary bricks of the metabolic network, based on reactions / genes / compounds found in BioCyc,
  - to manipulate and visualize metabolic graphs (Cytoscape)
  - also : export gene-reaction correspondances

- **Link Biocyc to other DBs, develop BioCyc-related algorithms**
  - Genoscope's LIMS use BioCyc compounds repository as a language of media representation,
  - Protein complexes comparative inference

- **Modify or implement new housekeeping functions for a PGDB**
  - add / update data from a given annotation database (e.g. multifun, cellular localization of genes products, pubmed links...)
  - delete / annotate pathways as part of curation process
    - keep that information even after annotation or BioCyc updates

Ongoing & future work

- Improve translation scheme between FRS and Java classes
  - Map complete inheritance hierarchy
  - Generate associations directly instead of using Ids
  - cardinality constraints

- Software platform dedicated to multi-CBM management
  - Basic constraint-based model management
  - Representation of sets of environments
  - Representation of perturbation sets, generation of corresponding models
  - Pluggable analysis modules (topological analyses, phenotype predictions…)

Open questions about OOCyc

- Would OOCyc be useful outside of Genoscope ?

- If so, what would be the primary uses ?
  - Java access to Biocyc entities
  - store PGDBs in relational databases
  - manipulate classes representing biological concept
  - …

- Packaging
  - jar files ?
  - eclipse project ?
  - eclipse plugin ?
  - Sourceforge project ?

Acknowledgements

Computational Biology Group

Pierre-Yves Bourguignon
Cyril Combe
Maxime Durot
Francois Le Fevre
Matthias Heinig
Benjamin Pinaud
Serge Smidtas

NeMo@genoscope.cns.fr

Metabolic Thesaurus

Alain Perret
Valérie Barbe
Gorges Cohen
Annett Kreimeyer
Nuria Fonknechten
Veronique De Bernardinis
Marcel Salanoubat
Jean Weissenbach

IT group

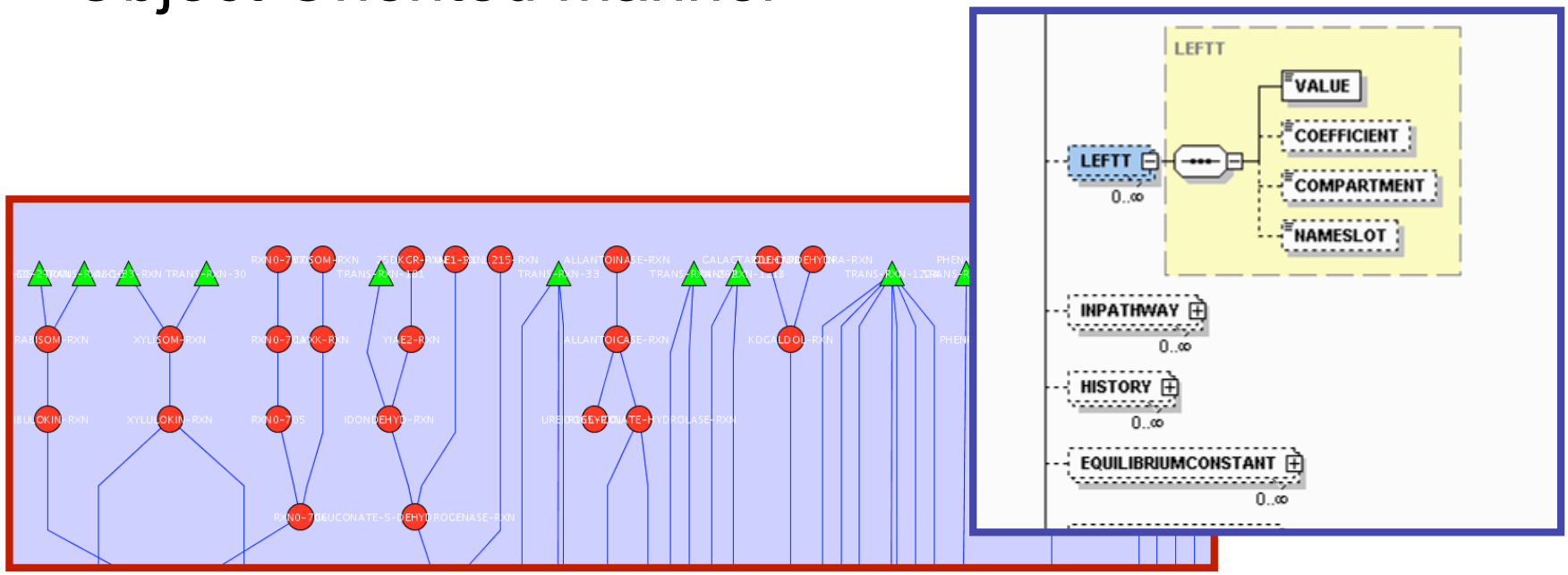
Ali Boudani
Laurent Sainte-Marthe
Claude Scarpelli

AGC
David Vallenet
Claudine Médigue

SRI
Peter Karp
& group

OoCyc:

the development of OoCyc, an additional API to query, manipulate BioCyc information in an Object Oriented manner

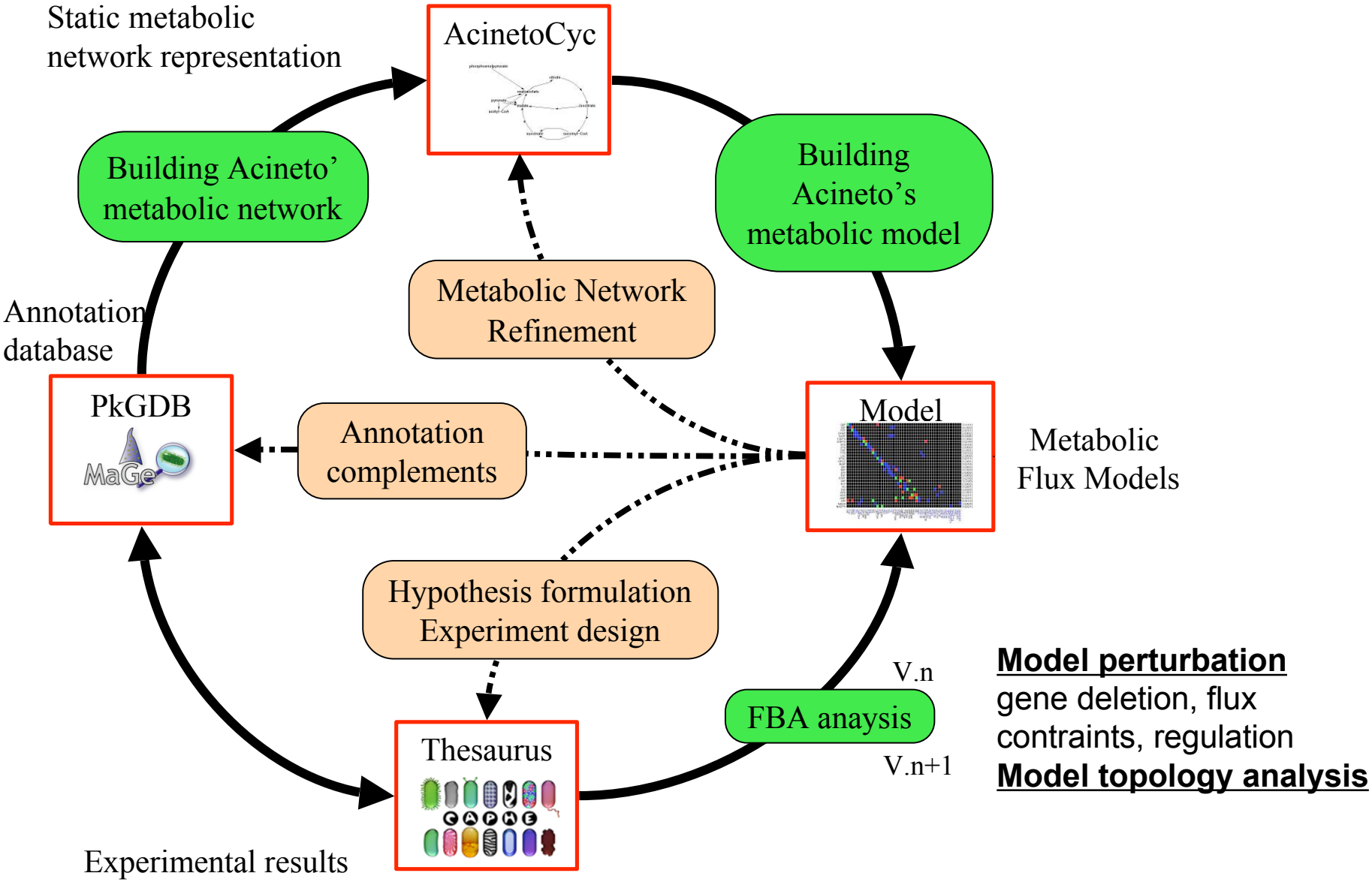


References

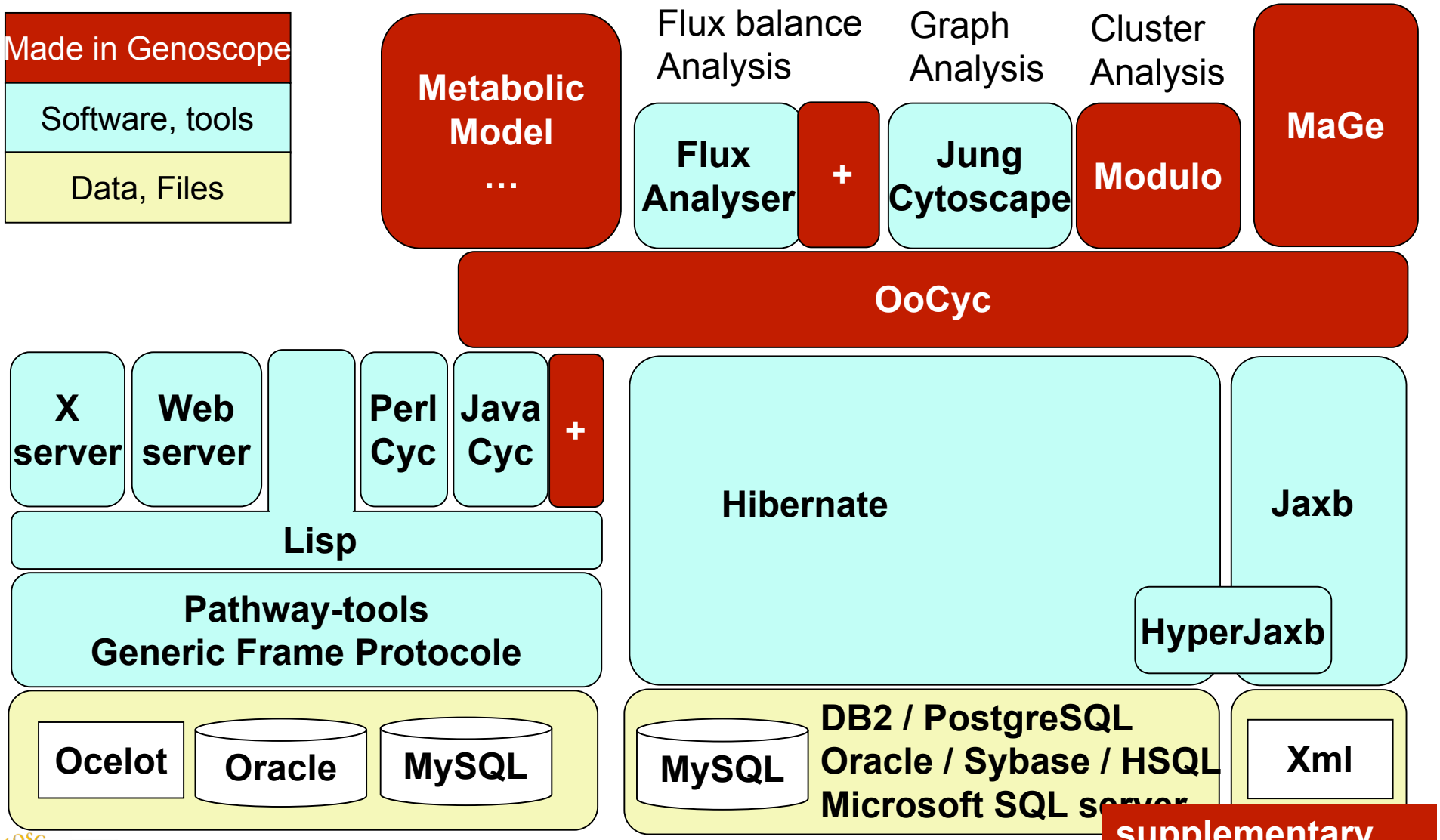
- **Flux balance analysis**
  - •Metabolic modelling of microbial strains *in silico* ; Covert et al ; Trends in Biochemical Sciences 2001, 26(3):179-186.
  - •Genome-Scale Metabolic Model of *Helicobacter pylori* 26695 ; Schilling et al ; Journal of Bacteriology 2002, 184(16):4582-459
  - •Metabolic pathways in the post-genome era ; Papin et al ; Trends in Biochemical Sciences 2003, 28:250-258.
  - •Theory for the Systemic Definition of Metabolic Pathways and their use in Interpreting Metabolic Function from a Pathway-Oriented Perspective ; Schilling et al ; J. theor. Biol. 2000, 203:229-248.
  - •Assessment of Metabolic capabilities of *Haemophilus influenzae* Rd through a genome-scale pathway analysis ; Schilling et al ; J. theor. Biol. 2000, 203:249-283.
  - •Metabolic network structure determines key aspects of functionality and regulation ; Stelling et al ; Nature 2002, 420:190-193.
  - •Characterizing the Metabolic Phenotype: A Phenotype Phase Plane Analysis ; Edwards et al ; Biotechnology and Bioengineering 2002, 77:27-36.
  - •*In silico* predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data ; Edwards et al ; Nature 2001, 19:125-130.

- **Biocyc /MetaCyc /EcoCyc**
  - •Karp P.D., Arnaud M., Collado-Vides J., Ingraham J., Paulsen I.T., Saier M.H. Jr. (2004). "The E. coli EcoCyc Database: No Longer Just a Metabolic Pathway Database." ASM News 70(1): 25-30.
  - •Cynthia J. Krieger, and Peter D. Karp, MetaCyc: a multiorganism database of metabolic pathways and enzymes, Nucleic Acids Research, 32(1):D438-42 2004.
  - • P. Karp, S. Paley, and P. Romero, « The Pathway Tools Software », , Bioinformatics 18:S225-32 2002.
  - • S. Paley and P. Karp, "Evaluation of computational metabolic-pathway predictions for H. pylori, "Bioinformatics 18(5):705-14 2002.
  - • Krummenacker M, Paley S, Mueller L, Yan T, Karp PD. , Abstract Querying and computing with BioCyc databases. Bioinformatics. 15;21(16):3454-5. 2005.

Our general strategy

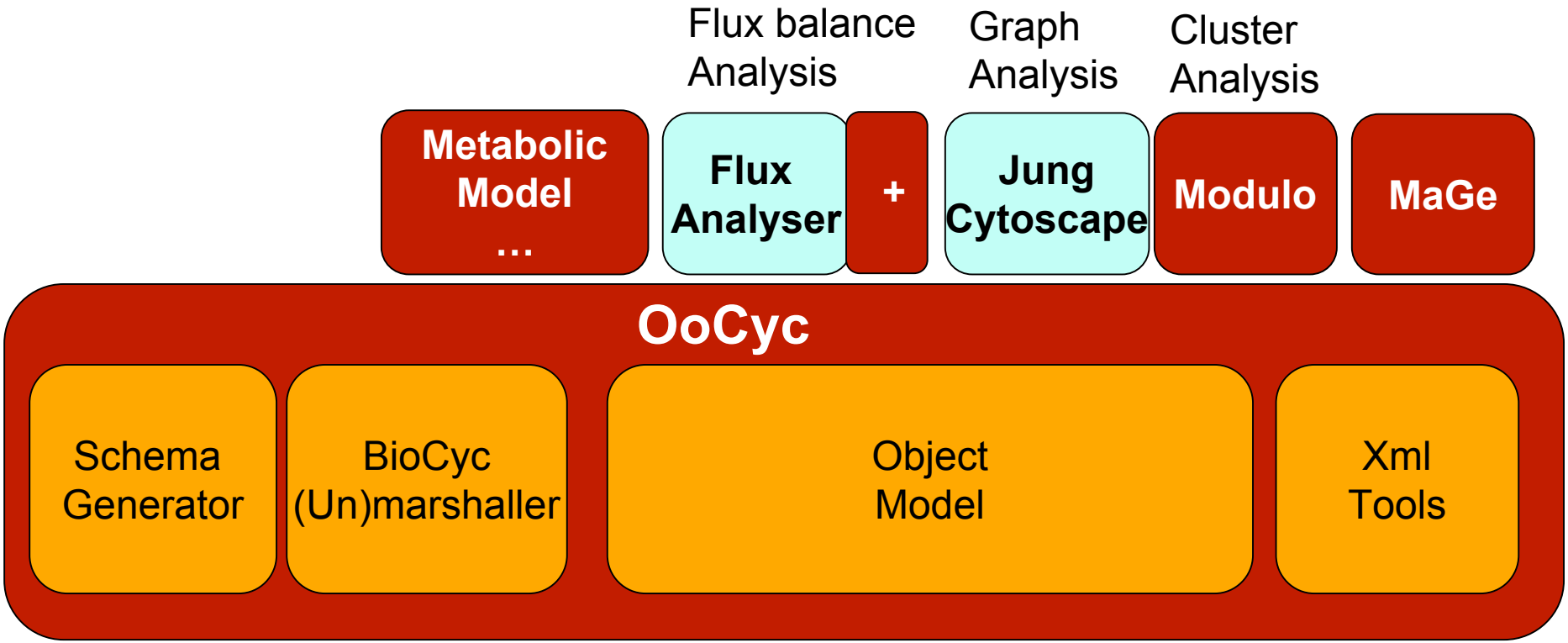


BioCyc and OoCyc's architecture



supplementary slides

OoCyc's architecture



Data, Files | Software, tools | Genoscope

True query in OQL

```
public List getAllErInhibitedBy(String org, String inhibitor) {
    try {
        Query hql = HibernateUtil.getSession(keySession).createQuery("select
    distinct er from
    fr.cns.genoscope.nemo.hypercyc.impl.ENZYMATICREACTIONSImpl er "+
                    " join er.INHIBITORSALLInternal inhibitors" +
                    " where " +
                    " er.ORGANISM.VALUE like :organism and " +
                    " :inhibitor = inhibitors.VALUE" +
                    " ");
        hql.setString("organism", org);
        hql.setString("inhibitor", inhibitor);
        return hql.list();
    } catch (HibernateException e) {
        e.printStackTrace();
    }
    return null;
}
```

Java API

```
package fr.cns.genoscope.nemo.hypercyc.cycs.acineto;

public class Acineto {
    public static void main(String[] args) {
        ENZYMATICREACTIONSDAO erdao = new ENZYMATICREACTIONSDAO("hypercyc");
        List r = erdao.getAllErInhibitedBy("ecoli","ATP");
        ENZYMATICREACTIONS er;
        for(int i=0; i < r.size(); i++){
            er = (ENZYMATICREACTIONS)r.get(i);
            System.out.println(

    er.getFRAME()+"\t"+((ENZYME)er.getENZYME().get(0)).getVALUE());
    }
}
```