Genome Mining for Enzyme Discovery: the role of bioinformatics

David Vallenet

LABGeM, CEA-Genoscope
Metabolic Genomics UMR8030

Genoscope and its research unit

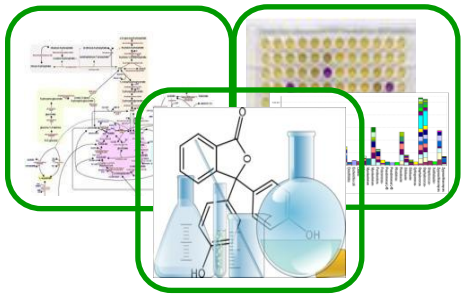
The **French National Sequencing Center** which is part of the France Génomique infrastructure (PIA 2011)

- Sequencing labs
- (bio-)informatics teams

FRANCE GÉNOMIQUE

A **fundamental research structure** « Metabolic Genomics »

⇒ extend the in silico sequence annotations using experimental approaches, in particular in the domain of enzymatic functions

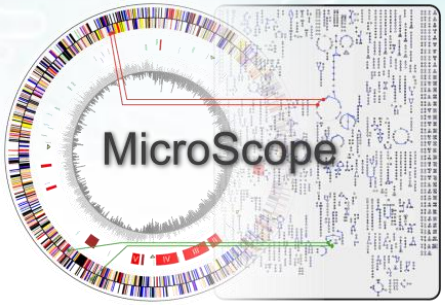


- Lab. of Genomics, Biochemistry, Organic Chemistry and Biocatalysis
- Lab. of Eukaryotic Genome Analysis (LAGE)

TARA OCEANS

- **Lab. of Bioinformatics Analysis in Genomics and Metabolism (LABGeM)**

LABGeM: Microbial Genomics



MicroScope, an integrated environment for:

- Automatic and expert genome annotation
- Comparative genomics (synteny, pan-genomes)
- Prediction of gene function and metabolic networks
- Gene expression (RNA-seq) and evolved strain (SNP) analysis

Data

- 10K genomes, 200 new genomes a month
- 3,900 user accounts
- 3,000 expert annotations a month

Service

- Academic and industrial genome analysis services
- Private instances for large strain collections
- 5-days training (500 trained users)

LABGeM: Microbial Metabolism

Genome analysis in a metabolic context

- Databases and tools for metabolic network reconstruction

SIB EMBL-EBI elixir Microbial biotechnology community

- Curation of several model species: *Escherichia coli*, *Bacillus subtilis*, *Pseudomonas putida*, *Acinetobacter baylyi*…

Discovery of new enzymatic activities

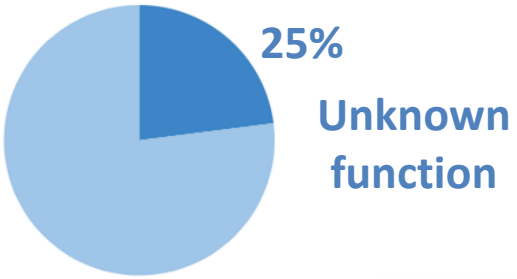
- By combining wet-lab experiments and bioinformatics
- Sequence analysis, genomic contexts, networks, structural analysis

Microbial Diversity in the Genomic Era

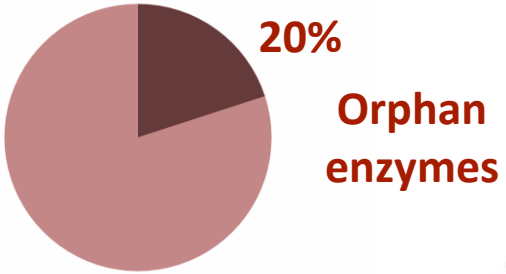
- The largest reservoir of enzymatic activities in the biosphere
- Incredible genetic diversity: millions of species, tiny fraction characterized

- Genomics: >100K genomes in databanks, >100M protein sequences
- Metagenomics: access to the diversity of uncultured species

>16000 protein domain families



>6000 enzymatic activities



Bioinformatics strategy

New enzyme families

Methods for the discovery of new enzyme families

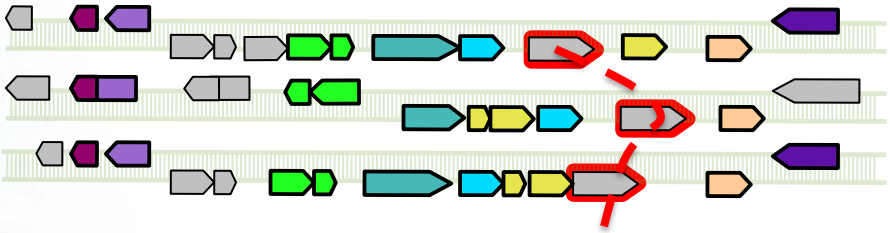
Strategy for the exploration of the functional diversity of enzyme families

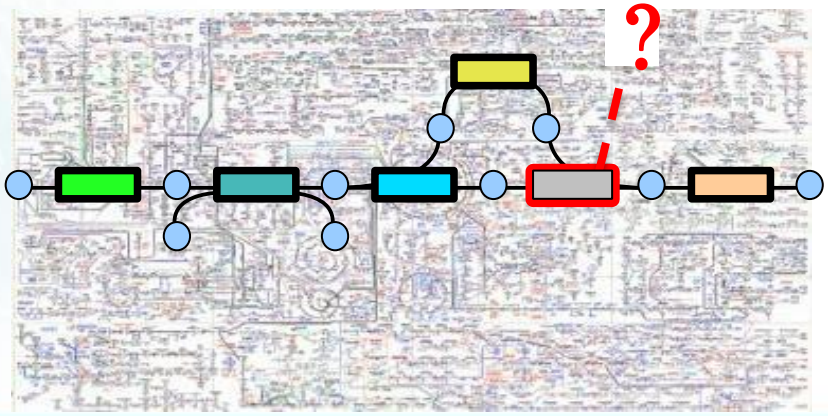
New metabolic contexts

Find new enzyme families

“Guilt-by-association” methods

- Based on comparative genomics (gene clustering, phylogenetic profiles, gene fusions)
- Based on post-genomic experiments (transcriptomics and proteomics data, gene essentiality)





Combine genomic and metabolic context

to link

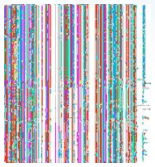
genes of unknown function
with
orphan reactions

‘Unknown’ proteins and ‘orphan’ enzymes: the missing half of the engineering parts list – and how to find it

Andrew D. Hanson[*,1], **Anne Pribat**[*], **Jeffrey C. Waller**[*], and **Valérie de Crécy-lagard**[†]
[*]Horticultural Sciences Department, University of Florida, Gainesville, FL 32611, U.S.A
[†]Microbiology and Cell Science Department, University of Florida, Gainesville, FL 32611, U.S.A

Explore the functional diversity of a family

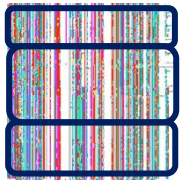
1. A protein family with one known function
2. Clustering in iso-functional groups and substrate screening on representatives
3. Study of the substrate promiscuity / specificity by 3D modeling
4. Genomic and metabolic context exploration



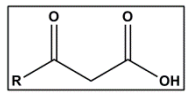
Multiple alignment

A +B <-> A' + B'

Generic reaction

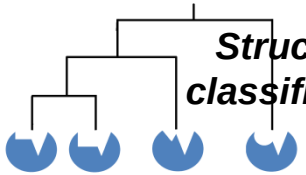


Clustering

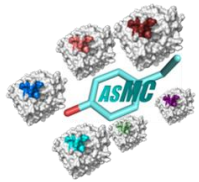


Potential substrates

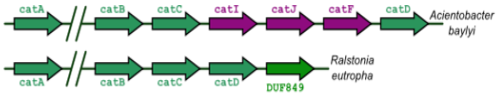
Screening



Structural classification



Identify key aa in the active site



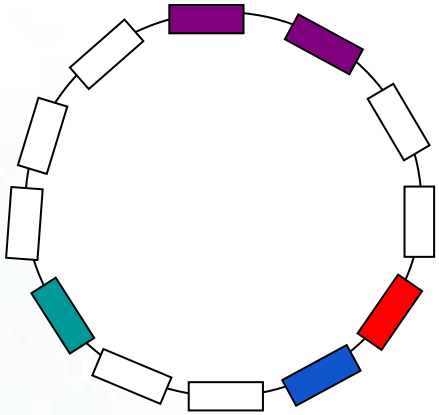
New metabolic pathways

ARTICLE
PUBLISHED ONLINE: 17 NOVEMBER 2013 | DOI: 10.1038/NCHEMBIO.1387
nature chemical biology

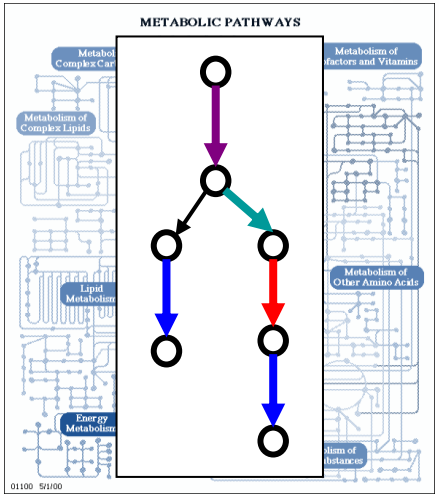
Revealing the hidden functional diversity of an enzyme family

Systems biology

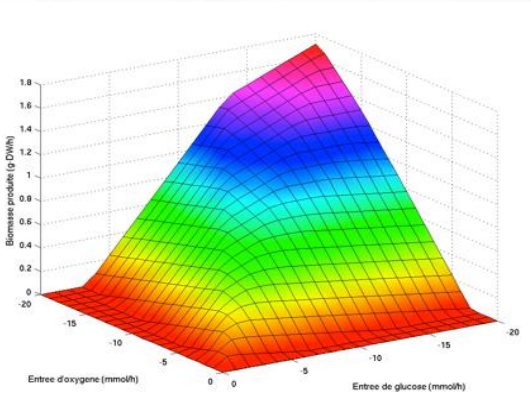
Genome analysis



Metabolic networks



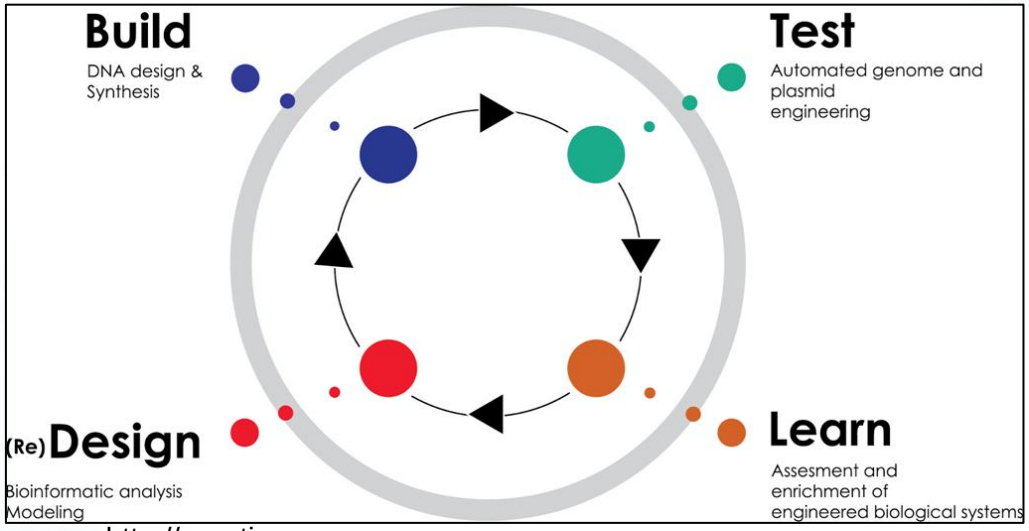
Metabolic models



Genome-scale Metabolic Models, a tool to:

- capture the knowledge of the metabolism of an organism and investigate its potential for chemical production
- identify key metabolic features such as growth yield, network robustness and gene essentiality

Synthetic biology



source: http://syncti.org

Computer-Aided Design
&
Machine Learning
+
experimental techniques
(DNA synthesis, genome construction, automated strain construction, high-throughput screening...)

- **Retrobiosynthesis:** chemo- and bio-informatics methods to define set of reactions and enzymes (genes) to build a desired target molecule from a central metabolite of a chassis organism
- **Computational enzyme design:** combine biochemical building blocks (amino acids, cofactors, coenzymes, etc.) to produce new catalysts

A review of computational tools for design and reconstruction of metabolic pathways

Lin Wang, Satyakam Dash, Chiam Yu Ng, Costas D. Maranas*

Angewandte Reviews

K. N. Houk et al.

Enzyme Design

DOI: 10.1002/anie.201204077

Computational Enzyme Design

*Gert Kiss, Nihan Çelebi-Ölçüm, Rocco Moretti, David Baker, and K. N. Houk**