Université d'Evry Val d'Essonne
Ecole Doctorale des “Génomes aux Organismes”

THESE
pour obtenir le titre de Docteur en Bioinformatique,
Biologie Structurale et Génomique

Présentée par
Adam Alexander Thil SMITH

Exploitation automatisée des contextes métabolique et génomique pour l'annotation fonctionelle des génomes prokaryotes

Automatically exploiting Genomic and Metabolic Contexts to aid the Functional Annotation of Prokaryote Genomes

Soutenance prévue le 03 février 2012 devant le jury composé de :

Présidente :	Florence d'ALCHE-BUC
Rapportrice :	Marie-France SAGOT
Rapporteur :	Antoine DANCHIN
Examinateur :	Alain VIARI
Examinateur :	Christos OUZOUNIS
Tuteur de thèse :	David VALLENET
Directrice de thèse :	Claudine MEDIGUE

Travail réalisé au sein du Laboratoire d'Analyses Bioinformatiques pour la Génomique et le Métabolisme, CEA/DSV/IG/CNS, UMR 8030 “génomique métabolique”

Nota bene: thèse rédigée en anglais

Table of Contents

List of Illustrations

I. Preamble

I.A. Thanks

First and foremost, I would like to thank Florence d'ALCHE-BUC for having graciously accepted to stand as president in my thesis jury. I am also very grateful to Antoine DANCHIN and Marie-France SAGOT for taking on the roles of thesis manuscript reviewers. I thank Christos OUZOUNIS and Alain VIARI for participating in my jury alongside them. Finally, I wish to take my hat off to Claudine MEDIGUE for having taken me on this thesis project, and to David VALLENET for having tutored me and put up with me over these past 3 years!

I also wish to present my fond thanks to all the members of the LABGeM team, past and present, with whom I have lived, worked and joked, and who made the difficult moments more fun to get through. Work-wise, I am especially indebted to Damien MORNICO and François LEFEVRE for helping me with Java- and MetaCyc-related topics, to Karine BASTARD for her invaluable interactions all along the the BKACE project, to Zoé ROUY for keeping me in check, and to David ROCHE for having accepted to look after Fudge every time a holiday became necessary!

I would also like to thank some members from other teams at the Genoscope with whom I enjoyed working: Valérie BARBE, whose good moods and help with genome sequencing was essential to this manuscript; Véronique DE BERARDINIS, Marcel SALANOUBAT, Annette KREIMEYER, Alain PERRET and Christine PELLE, the “first floor” biochemists without whom none of the experimental work presented in this manuscript would have been possible.

I am grateful to Muriel TESSI, secretary to a whole third-floor-full of computational biologists and computer scientists - no easy task! - who was always there with a smile to help out for all my Genoscope-related administrative tasks. You are missed here!

Though it is not common to do so, I wish to thank the people of the administrative scholarship service, and of the Ecole Doctorale “des Genomes aux Organismes”, from the Univeristy of Evry. Indeed, having been able to see how complicated singing up to a university for a thesis can be thanks to the life experiences of others, I am truly grateful to these people, and especially to Florence HAMON, for having made the administrative aspects of a thesis so comparatively *simple* at the University of Evry. May other universities take example on you!

Obviously, I do wish to thank my parents for their support and counsel throughout my thesis, that they even helped proofread, no meagre task! Last but not least, I am very grateful to Tiffany SMITH (PhD!), my wonderful wife (wedded during our theses, indeed!), for her care, support and love during these three eventful years, without whom all of this would have been meaningless. Thank you!!!

I.B. Notes

Terms written in **bold** font are those which are of particular relevance to their host paragraph, and are designed as **eye-catchers** for quick reading. Terms written in *italic* and that were not required to be indicate *emphasis*.

All articles inserted into this manuscript *will not count towards page numbering*; furthermore, their bibliographies will be separate.

All numbers concerning data were established during August, 2011, unless otherwise indicated.

I.C. Summary

The projects I have participated in during my thesis all concern the functional annotation of prokaryote genomes with the help of bioinformatics tools. Indeed, the masses of data produced by high-throughput sequencing cannot be processed manually as was done in the past. Instead, the computational use of mathematical and statistical models has become an absolute necessity for storing and handling the data, for extracting useful knowledge from it or for copying it to newly-sequenced genomes.

The general objective of my thesis work was to develop new bioinformatics methods for the (semi-)automated functional annotation of prokaryote genomes. In this manuscript, I first present the current scientific context around this objective, from genome sequencing to metabolism. I then present the two main projects I have worked on, that were both dedicated to the functional annotation of prokaryote organisms.

The first of these involved the development of the **CanOE strategy**, which focuses on finding candidate genes for sequence-orphan enzymes, integrating results across all available prokaryote genomes, and making these propositions available to the scientific community. I developed a generic method for building metabolic networks using publicly available metabolic databases, as well as a programmatic wrapper for the CCCPart algorithm, which I use to locate genomic metabolons in prokaryote genomes. These metabolons serve as a basis for finding candidate genes for the sequence-orphan enzymes, and results are considered across over 1,000 genomes from the MicroScope bioinformatics platform in order to evaluate confidence scores. Benchmarking experiments and manual bioanalysis of several results revealed that the strategy was indeed useful in several ways. I finally worked on integrating this bioinformatics tool into the MicroScope platform in order to make it available to the scientific community.

In a second project called the "**BKACE project**", I computationally helped explore the functional space of a recently and partially characterised enzyme family (DUF849), which led to the creation of novel functional annotations for its member genes. By combining sequence- and context-based information, I divided the family into *à priori* iso-functional sub-families, and using chemical substructure searches proposed some potential substrates for the enzymes from them. Our collaborators carried out a high-throughput biochemical assay of over 20 substrates across almost 200 DUF849 proteins; I then worked on analysing the data and graphically representing it. All in all, many novel enzymatic activities were demonstrated to be catalysed by the sub-families of DUF849, and we hope to be able to adapt our analysis strategy to other gene families.

I also participated in other projects, using my knowledge of **multivariate statistical methods** on metabolic data sets that my co-workers were analysing. These analyses were intended to be useful in generating high-level views of the metabolic capacities of the genomes under study, which were instrumental in perfecting their functional annotations. Several of these projects led to publications, of which two are included here as annexes.

Finally, I tie all these works together under the banner of my thesis subject, discussing what contributions to the field they may bring, in which ways they can be combined, and some perspectives of improvements that could be made to them.

I.D. Résumé

Nota bene : Ce résumé correspond approximativement à une traduction du chapitre d'introduction (Chapitre II) plutôt que du résumé anglais, puisque ce premier se devait d'être plus conséquent, à la demande de l'école doctorale.

Tandis que les technologies de séquençage deviennent plus puissantes, rapides et moins onéreuses, la validation expérimentale de l'ensemble des fonctions des gènes nouvellement séquencés est passée d'extrêmement coûteuse à tout simplement inimaginable. Depuis plus de vingt ans, le domaine de la bioinformatique s'est développée, en proposant des outils capables de guider les expérimentations ou de les remplacer. La technique de base la plus largement utilisée repose sur l'identification de liens évolutifs entre gènes en analysant la similarité entre leurs séquences (au niveau des nucléotides ou des acides aminés), qui peuvent alors être utilisés pour transférer des annotations fonctionnelles des uns vers les autres. Cette technique a été incroyablement utile pour propager les connaissances établies sur un petit nombre d'organismes modèles (c'est-à-dire fortement étudiés) vers des organismes nouvellement séquencés. Cependant, il a été montré que ce genre de transfert était source d'erreurs [1].

D'autres techniques ont été développées pour contourner ces inconvénients. L'utilisation d'informations contextuelles, plutôt que les séquences brutes, pour la prédiction de fonctions de gènes est l'un des axes principaux de recherche en ce sens. Le sujet de ma thèse s'inscrit directement dans cette problématique. Plus précisément, elle était dédiée au développement d'outils ou de stratégies bioinformatiques exploitant de l'information de contextes génomiques et/ou métaboliques afin de générer des annotations fonctionnelles potentielles. Le projet principal de ma thèse avait un objectif plus ciblé encore.

En effet, environ 27% des activités enzymatiques définies par le « International Union of Biochemistry and Molecular Biology » (IUBMB) sont encore aujourd'hui des activités orphelines de séquence, c'est-à-dire que malgré une connaissance biochimique de leur existence, aucun gène codant ni protéine n'a été identifié comme acteur de leur catalyse. Ceci empêche nécessairement l'utilisation des méthodes à base de similarité de séquence pour leur trouver des gènes candidats. De plus, le prix de la réalisation de l'ensemble des tests biochimiques nécessaires à l'identification d'un gène codant parmi tous les gènes de fonction inconnue dans un seul organisme serait rédhibitoire. Il est donc impératif d'utiliser des méthodes bioinformatiques afin de réduire le champ des possibles en trouvant un nombre limité de gènes candidats pour ces activités enzymatiques orphelines.

Certains travaux ont déjà été entrepris afin de résoudre -au moins partiellement- le problème des activités orphelines [2,3]. Cependant, bien peu de méthodes capables de proposer des gènes candidats pour les activités orphelines ont été conçues. Le projet principal de ma thèse consistait à développer une stratégie capable de répondre à ce besoin. Le premier travail que j'ai réalisé est la stratégie CanOE (fishing Candidate genes for Orphan Enzymes) qui a été mis en place au sein de la plate-forme de bioinformatique MicroScope [4]. J'ai développé une méthodologie de construction générique de réseaux métaboliques à partir de bases de données publiques, et adapté l'algorithme CCCPart [5] afin de localiser des métabolons génomiques dans les génomes prokaryotes de MicroScope. Ces métabolons servent de base pour la proposition de gènes candidats pour les activités enzymatiques orphelines de séquence. Ces résultats sont alors intégrés sur plus de 1000 génomes afin d'établir des scores de confiance. Une procédure de validation informatique, ainsi que la bioanalyse manuelle des résultats, ont montré que CanOE était utile à plus d'un égard. Finalement, j'ai travaillé à l'intégration de cet outil parmi ceux de la plateforme MicroScope afin de rendre les résultats accessibles à la communauté scientifique.

Il existe un problème miroir à celui des activités orphelines de séquence : celui des protéines de fonction inconnue. De nombreux travaux ont été entrepris pour tenter d'associer des annotations fonctionnelles plus ou moins précises à ces protéines et à leurs gènes codants (*c.à.d.* des méthodes réalisant du transfert d'annotation sur base de similarité de séquence et des approches plus complexes comme l'utilisation experte de plate-formes d'annotation telles que MicroScope [6,4]). Un projet collaboratif appelé « projet BKACE » (pour « beta-keto acid cleavage activity ») a été initié suite à la découverte, au Genoscope, d'une nouvelle activité enzymatique pour quelques membres d'une famille de gènes (dite « DUF849 ») de fonction jusqu'alors inconnue. L'objectif de ce projet était l'exploration de l'espace fonctionnel de cette famille. Mon rôle dans ce projet a été de développer une méthodologie exploitant des informations contextuelles et de séquence afin de 1) découvrir des fonctions enzymatiques potentielles au sein de la famille et de 2) réduire le nombre de protéines à tester en séparant la famille en sous-familles *à priori* iso-fonctionnelles. Ces deux points ont servi à guider les expérimentations biochimiques menées par nos collaborateurs, qui ont testé au final plus de 20 substrats contre presque 200 protéines. J'ai réalisé l'analyse statistique de ces résultats et les ai préparés pour représentation graphique. Au final, plusieurs nouvelles activités enzymatiques ont pu être associées à des sous-familles de DUF849, et nous espérons pouvoir adapter la stratégie développée ici à d'autres familles enzymatiques.

J'ai également participé à d'autres projets, où j'ai appliqué mes connaissances en méthodes d'analyse statistique multivariée à l'étude de données métaboliques sur lesquelles travaillaient mes collègues. Ces analyses ont servi à générer des visions abstraites de haut niveau des capacités métaboliques des génomes étudiés, utiles à leur annotation fonctionnelle manuelle. Plusieurs de ces projets ont mené à des publications, et deux de celles-ci sont données en annexe.

Le premier projet visait particulièrement à trouver des gènes candidats pour des activités enzymatiques orphelines de séquence ; le second à aider à l'exploration de l'espace fonctionnel d'une famille de gènes à la fonction inconnue ; les secondaires à établir des connaissances utiles à l'annotation manuelle. En somme, les travaux effectués pendant cette thèse avaient pour objectif l'utilisation et la création de méthodologies bioinformatiques exploitant des informations contextuelles afin de générer des annotations fonctionnelles potentielles entre des gènes issus de génomes prokaryotes et des activités enzymatiques.

Dans ce manuscrit, je présente en premier lieu la source principale des données exploitées par la bioinformatique : le séquençage de génomes et leur analyse, ainsi que quelques projets d'intérêt de séquençage en cours. Je présente alors le domaine de l'annotation fonctionnelle de genomes prokaryotes, les plate-formes qui existent pour ce faire, et les stratégies couramment utilisées (Chapitre III). Je détaille ce qu'est effectivement l'annotation fonctionnelle, ses méthodes, et en propose une classification (Chapitre IV). Je présente alors le métabolisme des prokaryotes, puisque dans ces travaux, il s'agit surtout de trouver des fonctions métaboliques aux gènes (Chapitre V). Je conclus mon état de l'art par le détail du concept d'activité enzymatique orpheline de séquence, et les solutions existantes pour y remédier (Chapitre VI). Je détaille ensuite mes travaux (le projet CanOE dans le Chapitre VII, le projet BKACE dans le Chapitre VIII, et des projets secondaires dans le Chapitre IX). Je développe des perspectives globales dans le Chapitre X. Les annexes et la bibliographie figureront dans les Chapitres XI et XII.

I.E. List of abbreviations

Abbreviation	Meaning
ASMC	Active Site Modelling and Clustering, a clustering method based on identifying key amino acids in enzyme active sites.
ATP	Adenosine Tri-Phosphate, a common energetic substrate heavily used in cell metabolism
BAC	Bacterial artificial chromosome
BKACE	Beta-keto acid cleaving enzyme/enzymatic activity
BLAST	Basic Local Alignment Search Tool, a bioinformatics tool for locating similar regions between protein or nucleic sequences
bp	base pair(s)
CA	Correspondence Analysis, a type of factorial analysis
CanOE	Fishing Candidate Genes for Orphan Enzymes
CDS/fCDS	Coding Sequence / fragment of a CDS
CoA	Coenzyme A, a common cofactor in enzymatic reactions
COG	Cluster of Orthologous Genes
ChEBI	Chemical entities of biological interest database, hosted by the EBI (see below)
DNA	Deoxyribonucleic acid, generic name for the helicoidal, double-strand assembly of deoxyribonucelotides that form a cell's genome.
DDBJ	DNA Data Bank of Japan
EBI	European Bioinformatics Institute
EC number	Enzyme Commission number, describing an enzymatic activity
ENA	European Nucleotide Archive
EU	European Union
FA	Factorial Analysis
GO	Genomic Object, an object modelling an annotation of a stretch of DNA in the MicroScope platform.
GOA	Genomic Object Annotation, an object I propose in my alternative MicroScope data model
GO term	gene ontology term
HGT	horizontal gene transfer
HMFA	Hierarchical Multiple Factorial Analysis
ID	Generic abbreviation for “identifier”. SQL tables may have “ID” (uppercase) in their name. SQL table columns may have “id” (lowercase) in their names.
IUBMB	International Union of Biochemistry and Molecular Biology
IUPAC	International Union of Pure and Applied Chemistry
KEGG	Kyoto Encyclopedia of Genes and Genomes, actually a more metabolism-centred bioinformatics resource
LABGeM	Laboratoire d'Analyses Bioinformatiques de Génomique et du Métabolisme Genoscope team
LABIS	Laboratoire d’Analyses Bioinformatiques des Séquences Genoscope team (now merged with LABGeM)
LCAB	Laboratoire de Clonage et de criblage des Activités de Bioconversion Genoscope team
LCOB	Laboratoire de Chimie Organique et Biocatalyse Genoscope team
LGBM	Laboratoire de Génomique et de Biochimie du Métabolisme Genoscope team
LDA	Linear Discriminant Analysis

Abbreviation	Meaning
MaGe	Magnifying Genomes, graphical interface for the MicroScope platform.
Mb/Gb	Mega bases, giga bases, units of quantities of deoxyribonucleotides in a sequence.
MCA	Multiple Correspondence Analysis
MFA	Multiple Factorial Analysis
MPL	Minimum Path Length
MSA	Multiple Sequence Alignment
NCBI	National Center for Biotechnology Information (USA)
NGS	Next Generation Sequencing
PkGDB	Prokaryotic Genome Database, data management system of the MicroScope platform
PLoS	Public Library of Science
PPI	Protein-protein interaction (network, data...)
RNA	Ribonucleic acid, generic name for a single-strand assembly of ribonucelotides
RPAIR	Reaction Pair, a KEGG concept for modelling chemical group transfers in metabolic reactions.
PHF	Pathway Hole Filler tool. PHF-GC refers to PHF - Genomic Context, a more recent version of PHF.
PIR	Protein Information Resource
PubMed	Published articles and citations for biomedical literature from MEDLINE, life science journals, and online books
SIB	Swiss Institute of Bioinformatics
SNAP	Similarity-neighborhood approach for finding functionally related genes
SQL	Structured Query Language
STRING	Search Tool for the Retrieval of Interacting Genes
SVM	Support Vector Machine, a family of statistical supervised learning methodologies
UK	United Kingdom
USA	United States of America
WIT	the "What Is There" integrated system for high-throughput genome sequence analysis and metabolic reconstruction

II. Introduction

As sequencing technology becomes more powerful, cheaper and accessible, experimentally validating the functions of all the newly sequenced genes has gradually gone from extremely resource-consuming to downright unimaginable. Over the past two decades, **bioinformatics** (also known as computational biology[1]) tools have been developed in order to help guide experiments, or to replace them altogether. The most widely used technique involves detecting evolutive relationships between genes by exploiting similarity between their sequences, and on this basis, transferring the **functional annotation** of one to the other. Incredibly useful for propagating knowledge from well-studied organisms to newly sequenced ones, tools based on this technique have been shown on one hand to be error prone, and on the other hand to be approaching a usefulness threshold, as most of the widespread conserved genes are now thought to be known [1]. Other techniques have been developed in order to resolve or circumnavigate these drawbacks. The use of **contextual information** (rather than sequence information) is one of the pursued avenues of research in this direction. The subject of my thesis was designed to participate in this research. More specifically, it was dedicated to the development of bioinformatics tools or strategies using genomic and metabolic contextual information in order to assist with the functional annotation of prokaryote genomes. It did, however, have a more precise goal than general functional annotation.

As of today, roughly 27% of all enzymatic activities documented by the International Union of Biochemistry and Molecular Biology (IUBMB) are **sequence-orphan activities** (orphan enzymes for short), meaning that despite biochemical knowledge of the activity, no coding gene sequence nor protein sequence has ever been established for the catalysing enzyme. This obviously proscribes the use of all sequence-based transfer techniques mentioned above. Extensively carrying out wet-lab surveys for identifying, for each sequence-orphan activity, coding genes amongst all un-annotated genes would also be prohibitively resource-consuming. Bioinformatics methods are thus a prerequisite to help guide experiments and reduce the gene and function spaces to explore.

However, although some high-level work has been done to address the orphan enzyme problem [2,3], relatively few methods have been designed to **propose candidate genes for orphan enzymes**. The main focus of my thesis was to develop a methodology capable of doing specifically this. The first work presented in this manuscript is thus **the CanOE** strategy (finding Candidate

[1] These terms can currently be considered as interchangeable, though efforts to define each separately are under way.

genes for Orphan Enzymes), that we put together as a local solution to this problem.

There exists a mirror problem to that of orphan enzymes: that of proteins of unknown function. Much more work has been done on trying to associate more or less detailed functions to these proteins, from the previously-mentioned simple homology-inferring methods to using annotation platform tools such as those in MicroScope [6,4]. Stemming from an initial discovery of a novel enzymatic activity associated to a protein family of previously unknown function, a collaborative project called the **BKACE project** (for beta-keto acid cleaving enzyme) was initiated at the Genoscope. This project had as objective the **exploration of the functional space of the family of proteins of unknown function**. I worked on this project, using contextual information to attempt to discover alternate protein functions, and to help guide the biochemical assays.

To summarise, the tasks in this thesis were carried out with the objectives of creating new bioinformatics methodologies exploiting contextual information to generate functional annotations between genes from prokaryote genomes and enzymatic activities. The first project specifically aimed to help find candidate genes for orphan enzymes; the second to help explore the functional space of a newly discovered enzyme family.

In this manuscript, I shall first go over the main source of data which bioinformatics tools process: genome sequencing and analysis, along with several noteworthy illustrative projects, before presenting the field of prokaryote genome annotation, its platforms and strategies (Chapter III). I shall detail what functional annotation is, and shall propose a new classification of annotation methodologies (Chapter IV). It will then be necessary to discuss prokaryote metabolism, as metabolic reactions are one of the types of function a gene may be annotated with (Chapter V). I shall conclude my state of the art by presenting the concept of orphan enzyme and the solutions that have been proposed to find parent genes for them (Chapter VI). I shall then be able to present and discuss the projects I have worked on during this thesis (CanOE in Chapter VII, BKACE in Chapter VIII, and other smaller projects in Chapter IX). Finally, a general discussion will attempt to place all of these works under the banner of my general thesis subject and will mention future developments and perspectives on them (Chapter X). Bibliography and annexes will be consigned to Chapters XI and XII.

III. Microbial Genomes and their Analysis

The availability of whole genome sequences for an increasing number of organisms has, amongst others, spurred the development of the heterogeneous field of bioinformatics along with our comprehension of Life itself. As is often in science, though, it has generated at least as many questions as it has answered, ensuring that bioinformatics remains an important emergent scientific domain.

In this chapter, I shall give an overview of how whole genome sequences came about, before giving a few perspectives in the shape of noteworthy sequencing projects that are currently underway or planned for the near future. I shall also point out a few sequence data resources that have become central to bioinformatics today. A word will be said about metagenomics, as one of my side projects involved metagenome analysis. Finally, I shall present the tools which are used to transform all this sequence data into usable, biologically-relevant information: annotation platforms.

III.A. Genome sequencing and resources

III.A.1. History and evolution of biological molecule sequencing

III.A.1.a. Birth of sequencing

As early as the late 19th century, much scientific attention was focused on determining the composition and structure of biological molecules. Proteins in particular were under close scrutiny, as they were hypothesised to be associated with the basic functions of life, or even to be physical support for heredity.

Determining protein sequences first became possible after the development of an experimental protocol based on protein fragment end group identification. This approach was imagined in Frederick Sanger's laboratory in 1949 [7–10], and led to the discovery of the amino-acid sequence of insulin (a peptide hormone with two subunits totalling 51 amino acids in humans) after 10 years of effort. Both P. Edman's protocol, developed in 1950 and based on a stepwise protein degradation approach, and A. Maxam and W. Gilbert's protocol, developed in 1972, were also very popular [11]. In 1953, Watson and Crick discovered the sequence-like molecular structure of nucleic acids [12–14] and postulated the “sequence hypothesis”, which stated that a) the biological specificity of a nucleic acid segment is expressed solely by its sequence in nucleotide bases (adenine, cytosine, thymine and guanine), that b) nucleic acid sequences are a simple code representing the amino acid sequence of a particular protein, and finally c) that these proteins are the building blocks the organism requires to exist and live.

Evidence supporting this informational rather than biochemical view was uncovered in the following decades[2]. Sanger was affected by Crick's ideas, and orientated his research towards directly sequencing nucleic acids rather than proteins. The year 1977 saw the birth of both **Sanger** and **Maxam-Gilbert DNA sequencing methods**. The latter method depended on cutting radioactive DNA molecules (maximum length: 250 nucleotides) at a random base of a given type (not exactly equivalent to each nucleotide type, but combinations thereof); electrophoresis of the fragments of different lengths obtained for cuts of each of 4 types led to a profile that could be interpreted in order to reconstitute the original sequence.

The Sanger method was based on synthesising DNA in presence of a mix of deoxyribonucleotides (dNTP) and dideoxyribonucleotides (ddNTP) that blocked the polymerase reaction. DNA synthesis primers, and later the ddNTPs themselves, were radioactively marked in order to render the synthesised DNA fragments ("**reads**") easily detectable. Electrophoresis of the fragments obtained using different mixtures of each of the 4 types of ddNTP generated interpretable profiles for reconstituting the final read sequence, which at the time was of roughly 250 nucleotides. The larger sequence length, along with the lesser toxicity of used reactants and ease of use, led the Sanger method to finally being preferred over the Maxam-Gilbert method.

Since, the Sanger method has been refined and automated. Two main developments are of note. The first is the use of fluorochromes rather than radioactive elements to mark the primers or the ddNTPs, which can be excited with a laser and read by an automatic optical apparatus. This allowed results to be automatically processed, rather than relying on manual inspection of an electrophoresis gel. This led to the development of the first partially-automated DNA sequencing protocol at L. Hood's laboratory in 1986 [15], followed by the marketing of the first gel-based automated sequencer, the ABI 370 by Applied Biosystems, in 1987.

The second development appeared at the end of the 1990's, and involved the replacement of gel electrophoresis by capillary electrophoresis, allowing much faster sequencing and higher throughput. The data produced by these sequencing techniques remained humanly analysable, and protocols derived from them are still used today.

[2] such as proof of the collinearity between DNA and protein amino-acid mutation positions, in 1964

Obviously, it was not possible to sequence an entire genome in one go using Sanger sequencing. Further more, multiple molecules of genomic DNA were required (for multiple fragments). In order to address these two problems, **DNA clonal banks** were created. To create a bank, the target genome was broken up into large, 200 kb fragments, and each fragment was inserted into a vector, creating **Bacterial Artificial Chromosomes** (BACs). Each BAC was then cultivated in pure colonies. These BACs generated sufficient quantities of target genome DNA fragments to be broken up in turn and inserted into "multicopy" plasmids. These plasmids could carry up to 5 kb of the target genome sequence, and a single vector could host one or two hundred copies of it. This provided enough genetic material for the final sequencing step by a Sanger method. The reads thus obtained then have to be **assembled** into a final sequence. This burden is placed on the assembly software, which relies on overlapping sequences, on established genetic maps of the target genome if available, an on BAC extremity sequencing. Finally, a **finishing** stage can be undertaken, consisting of targeted sequencing of previously "missed" or low-quality parts of the target genome.

III.A.1.b. The Human Genome Project

A major driving force behind the development of more advanced, cheaper, and faster DNA sequencing was the Human Genome Project. This scientific objective was officially launched in 1990 with the cartography of the human genome. However, the sheer volume of base pairs to be sequenced (approximately $3*10^9$) was beyond the reach of previous techniques, requiring experimental and computational improvements. Two competing efforts took up the challenge. The **International Human Genome Sequencing Consortium** was a publicly-funded effort coordinating laboratories worldwide, that used a BAC genetic map approach to separate the human genome into over 20 000 more manageable "bites". The total project cost roughly 3 billion US dollars. The Genoscope was part of this endeavour, and completed the sequence of human chromosome 14. Craig Venter, at the time president of **Celera Genomics**, led a privately-funded effort based on "shotgun sequencing", which claimed that a protocol using less BACs and relying on powerful sequence assembly software to piece together many more reads could outperform the older protocols. Over the years, the much slower public project published several high-quality draft versions of parts of the human genome, but building on the latter, Venter finally managed to publish his own, low-quality version of the complete human genome first [16]. The public project finished mere weeks later [17,18]. The speed and relative cost (a tenth of that of the public effort) led computer-heavy approaches such as Venter's to be the most popular in subsequent developments, and they are still used today, even though they require a more costly finishing phase to ensure a good final sequence quality.

III.A.1.c. Sequencing technologies today

Over the past ten years, the sequencing landscape has evolved into a fast-paced race towards data. Since 2004, new sequencing technologies that are not based on Sanger's protocol have appeared. For example, the necessity of creating libraries of bacterial clonal colonies has been reduced in some sequencing technologies by the use of PCR amplification within the sequencers. Three great families of techniques have emerged: those based on DNA synthesis, those based on hybridisation, and those based on single-molecule sequencing, though the latter are much more recent and only starting to be commercialised.

DNA synthesis-based sequencing techniques: **Pyrosequencing** is a sequencing technique which relies on step-by-step synthesis of a complementary DNA strand using “washes” of dNTP of known types. When a dNTP is successfully incorporated into the new strand, the liberation of diphosphate triggers a chemical cascade that emits light that can then be detected by a photosensor. **Illumina sequencing** is a stop-and-go synthesis-based sequencing technique. In each cycle, a mix of modified dNTPs are added, each type carrying a fluorochrome with a specific colour that blocks elongation, and all blocking the polymerase reaction. Under laser stimulation, the latest added dNTP can be “read” thanks to its colour. Then, a chemical reaction removes the fluorochrome and unblocks the new DNA strand, which is then ready for another cycle.

Hybridisation-based sequencing techniques: **SOLiD sequencing** (Sequencing by Oligo Ligation and Detection) is based on hybridising DNA fragments with pre-prepared oligonucleotides.

Currently available sequencers of each of the previous techniques are compared in the table below:

Technique	Make	Sequencer	Read length (bases)	Output	Run Time	Error rate	Cost per base
Sanger	Applied Biosystems	ABI 3730XL	700	96 reads/run	2 hours	Medium	High
Pyrosequencing	Roche	454 Genome Sequencer FLX Titanium	500	500 Mb/run	8 hours	High	Medium
Illumina	Illumina	Solexa HiSeq 2000	100	250 Gb/ flowcell	10 days	Low	Low
SOLiD	Applied Biosystems	5500xl series	60	20 Gb/day		Low	Low

b: bases (nucleotides)

All these techniques make up what is called "second generation sequencing" or "**next generation sequencing**" (NGS). NGS has rendered sequencing much more accessible to small laboratories, and new ways of using sequencing have emerged (such as RNA discovery, PCR amplicon analysis, expression level measuring "RNA-seq", metagenomics...) [19]. In 2009, the cost of sequencing a human genome was evaluated at 100,000$, 3,000 times less than Venter's first supposed success, bringing within reach previously unimaginable projects such as the 1,000 human genomes project (see part III.A.2.a). The increase in accessibility, in number of sequencing techniques, of use cases, and sheer data volume, have led to the development of a plethora of ways of dealing with the data, with each lab basically coming up with specific software designed to meet their needs. This obviously rendered a previously complex domain increasingly confused.

With sequencing technology available to all laboratories, it is thought that major centralised sequencing centres have outlived their prime mission and that they too, should evolve. New missions could include comprehensive development of sequence analysis software, establishment of data formats to ease data and competence transfer, and dispensing training courses in good sequencing practices and sequence analysis. This will hopefully help to harmonise the sequencing landscape and avoid misconceptions or false ideas [20]. In this social era, discussion forums such as SEQanswers [http://seqanswers.com/] will probably be informational hubs for sequencing-related questions.

Perhaps the "$1,000 genome" is no longer such an unimaginable goal [21] though the cost of sequence data storage and analysis is, for its part, on the rise [22].

III.A.1.d. Sequencing technologies for the near future

Several new technologies (also called “third generation technologies”) have emerged and are soon to be commercially available.

Pacific Biosciences Single Molecule Real Time (PacBio **SMRT**) sequencing relies on the synthesis of the complementary strand of a single DNA molecule in a tiny well. Individual types of dNTPs carry chromophores that are stimulated by laser when they are added to the DNA, after which they break off. It is thus possible to follow in real time the adding of successive dNTPs, and thus determine the sequence. The laser can be switched on or off, leading to strobe sequencing, particularly useful for reading longer and gapped sequences, as switching the laser off periodically reduces damage done to the DNA.

Ion Torrent sequencing also relies on a single complementary DNA strand being synthesised. The well is flushed in turn by different types of dNTP. When a dNTP is added to the strand, the release of hydrogen ions (H^+) is detected, allowing determination of the sequence.

PacBio SMRT is capable of generating large quantities of relatively long reads quickly (3Mb per hour, 1000 bp reads). It is, however, currently plagued by a low read quality. The Ion Torrent sequencer, however, is relatively small, portable, with a high turnover and read quality (<1% error rate), useful for quick desktop sequencing, but not for large genome sequencing projects (reads of 100-200 bp). As an example use case, in 2011, the speed and ease of Ion Torrent technology helped a team from the Lille Pasteur Institute to sequence the genome of a pathogenic *Escherichia coli* strain that had caused a widespread food intoxication across Europe in less than 3 days.

Even more technologies are emerging, and it is suggested to the interested reader to keep an eye out for these new developments, though all might not be successful.

III.A.2. Ongoing projects and resources

III.A.2.a. Genome projects

As sequencing technology and sequence assembly become faster, more reliable and cheaper, genome sequencing projects will become increasingly ambitious. Below, I describe three projects that plan to revolutionise bioinformatics and its applications in different ways.

Genomic Encyclopaedia for Bacteria and Archaea: There are currently over 2,000 bacterial and archaeal genomes available in public sequence data banks. However, sequenced genomes show heavy phylogenetic bias towards laboratory-cultivable organisms (estimated to represent less than 10% of existing microbes thanks to metagenomic analyses, see part III.A.2.b) [23,24]. Though sequencing more genomes from diversified phyla is not expected to result in many more biological discoveries due to "diminishing returns", it is argued that they are necessary for a) opening usual molecular biology study approaches to new organisms, b) obtaining a broader view of the tree of life, c) countering previously described sampling biases, d) assessing taxon diversity in this era of popular interest in biodiversity [24], e) improving gene/protein family detection (see part IV.C.2), f) winkling out novel biological discoveries, and more.

To this end, leading researchers have instigated projects like the Genomic Encyclopaedia for Bacteria and Archaea (GEBA) and the USA National Science Foundation project "**Assembling the Tree of Life**". The GEBA project is an international effort started in 2008 and that now covers roughly 100 recently-sequenced microbial genomes (estimated from a PubMed search of articles following the GEBA publication guidelines). The added value of increased and less biased phylogenetic coverage has already been shown [25].

The GEBA project is followed up by an interesting educational program created by the Joint Genome Institute (JGI) of the Department of Energy (DOE) of the USA, entitled "Interpret a Genome" [http://www.jgi.doe.gov/education/interpretagenome.html]. This program proposes that sequenced GEBA genomes be annotated in the context of biochemistry and bioinformatics undergraduate courses in universities across the world, and several genomes have already been annotated over past yearly sessions.

The 1,000 Genomes Project: The 1,000 Genomes Project [http://www.1000genomes.org/], launched in 2008, aims to completely sequence (with good coverage) the genomes of a large number of humans in order to establish the first comprehensive map of human genetic variability (sufficient for 95% coverage of all alleles[3] present in over 1% of the global population). This will obviously be of great scientific value to multiple domains of medicine, and should lead to interesting discoveries in the field of population genomics (which studies the evolution of genetic variations over multiple generations). It should also be able to complement results obtained from

[3] An allele is one of several variants of a gene (i.*e.* differing by one or several nucleotide polymorphisms) .

human metagenomics projects by shedding further light on the relationships between a human host genome and its harboured microbial communities (see next section III.A.2.b). With the development of cheaper, faster sequencing methods, this project is becoming less of a challenge and more of a research milestone. Indeed, pilot sub-projects have already been completed, leading to interesting results validating the approach [26].

The 10,000 Genome Project: The Beijing Genomics Institute (China) has organised the 10,000 Genome Project, which is similar to the Genomic Encyclopaedia of Bacteria and Archaea project, but specifically for microbes present in conventional environments, extreme environments, and human body samples, all geographically located in China. This will thus call on traditional genomics and metagenomics (see III.A.2.b). Archaea, bacteria, fungi, algae and viruses will be studied in the hope of uncovering biological particularities with industrial applications. A description of this project can be found at [http://www.genomics.cn/en/research.php?type=show&id=498].

Project to Annotate 1,000 Genomes: In 2003, the Fellowship for the Interpretation of Genomes (FIG) [www.thefig.info] initiated the Project to Annotate 1,000 Genomes, which had as primary objective to develop methodologies for accurate and high-throughput annotation of genomes from all domains of life, while still maintaining expert curation as an active part of the process in order to guarantee necessary precision and error correction. Though this is not strictly speaking a “genome” project (dedicated to annotation, not sequencing), it is still of relevance in the context of this thesis. Is namely formed the heart of the collaborative effort out of which the SEED was born [www.theseed.org] (see section III.B.3).

As announced, science has advanced from sequencing and studying single genomes to multiple genomes, which I shall now briefly describe.

III.A.2.b. Metagenomics and metagenome projects

Metagenome sequencing is an even younger science than single genome sequencing, and has opened several unexpected doors into the biology of micro-organism communities. Here, I shall briefly define what is metagenome sequencing, along with its generic protocol, before presenting other projects of interest in this expanding domain that hold high hopes for future bioinformatics developments.

A major fraction of all living micro-organisms currently evade biologists' attempts to cultivate them in isolated, laboratory-controlled conditions ("clonal cultures"). Studying their inner workings, their metabolism and interactions is thus rendered infinitely more complex than for species easier to cultivate, let alone sequencing their genomes. However, early sequencing attempts of non-clonal samples showed that biodiversity had been grossly underestimated by lab cultures [27], as many environments absolutely team with life, such as soil, digestive tracts and sea water.

Modern molecular biology techniques and micro-organism genomics have now opened a door to the workings of these evasive organisms. It is possible to sequence and study the genomes of many different, non-isolated organisms from a single environmental sample, a process referred to as metagenomics analysis [28]. This allows the study of micro-organisms without the biases introduced by laboratory limits [27]. The scaling up in quantities of sequenced DNA, and the mixture of original species that can be quite close, pose new computational problems for their analysis, and has spawned a fervent era of bioinformatics research [29].

Metagenomics studies all follow a same general procedure. The first step is, obviously, to retrieve a sample of which the metagenome is to be sequenced. This usually involves environmental sampling, where experimental protocols much be chosen and followed scrupulously to ensure that there is enough genetic material to detect all of the species in the sample (the number and distribution of which should be estimated), and that it is not contaminated by unwanted organisms (*e.g.* eukaryotes, small multicellular organisms, human cells...). The metagenome can then be constructed by extracting the DNA from the sample, generating reads of this DNA, assembling the reads into contigs (stretches of high-confidence DNA sequence), and then into scaffolds (even longer sequences supposed to belong to a single organism). One can attempt to use codon usage or sequence homology (see IV.B.1) to assign scaffolds to already-known or related taxa (identified by Operational Taxonomic Units, species distinctions in microbiology), a process known as "binning". Once ready, a metagenome can be annotated just as a set of single genomes might be, using automated methods and/or manual expertise (see III.B). Several analyses are available: biodiversity evaluation by estimating the number of species present and their distribution; reconstructing the metabolisms of each specie and underlining their ecological interactions, establishing the phylogeny of the species present, *etc*.

Many projects exist that aim to take advantage of what metagenomics has to offer in order to extend our comprehension of the living world that surrounds us. I shall present a select few below.

The Human Microbiome Project: It has long been known that a human body does not live in isolation in respect to other organisms, but instead is actually host to a plethora of living entities that co-exist symbiotically[4] with it; for example, the number of microbial cells in the human gastrointestinal tract far outnumber the number of cells in the human body [30,31]. With the advent of metagenomics it has become possible to study these complex populations. The Human Microbiome Project [https://commonfund.nih.gov/hmp/] intends to elucidate the workings of the complex microbial populations that inhabit human skin, gut, and other mucosal cavities, hopefully discovering how they interact with their host, and how perturbations of human microbiomes may result in diseases [http://en.wikipedia.org/wiki/Human_microbiome_project]. The **MetaHIT project** is a European subsection of this effort, that focuses on gut metagenomes, and in which the Genoscope is currently participating.

Ocean sequencing expeditions: Genome research centres worldwide have collaborated within two recent adventures into the little-explored biodiversity of the ocean. Both projects involved scientific teams manning an adapted sea vessel, sailing it around the world, sampling sea water at various depths, isolating the various micro-organisms thus captured, and performing basic biochemical experiments while waiting for the samples to be shipped off to sequencing centres at the next port of call. The first of these is **the Global Ocean Sampling Expedition** [http://www.jcvi.org/cms/research/projects/gos/] and was formed by the J. Craig Venter Institute (USA), while the second is the **Tara Oceans Expedition** [http://oceans.taraexpeditions.org/], a European initiative in which the Genoscope is participating, focusing primarily on marine plankton.

It is hoped that these endeavours will help complete our understanding of marine ecosystems, biodiversity, and of evolution in general. They should also contribute to the establishment of a marine cartography of geological, physico-chemical properties, as well as discovering biomarkers of marine pollution. They might also offer new opportunities for biotechnologically harnessing new agronomic, energetic, medical and cosmetic resources.

[4] The relationship can thus be commensal, mutual, parasitic, or any intermediate of these.

All these projects would be of little use without a system for exchanging sequence information and data, hence the birth of biological sequence resources.

III.A.2.c. Biological sequence resources

Sequenced gene and protein sequences, along with their known functions, have been consigned to specialized publicly-accessible databases and databanks, allowing researchers to retrieve at will any information they require for their work.

INSDs: One of the main sequence resources is the set of International Nucleotide Sequence Databases, of which specific mirrors are the NCBI's GenBank (USA), the DDBJ (Japan), and the EBI's ENA (Europe). The INSDs are maintained collectively by the International Sequence Database Collaboration [http://www.insdc.org/], in order to contain precisely the same data for a given release. They collect all known publicly-available nucleotide sequences (DNA & RNA) to which several annotations of different types are attached (genes, coding sequences, functional roles, located domains, bibliographic references, exon/intron mapping...). INSDs are designed to be primary resources for nucleotide sequence data.

RefSeq: The RefSeq collection [http://www.ncbi.nlm.nih.gov/RefSeq/] is maintained by the NCBI. It collects non-redundant DNA, RNA and protein sequences that are annotated with much more detail (generally of medical interest) than what is contained in the INSDs.

UniProt: UniProt [http://www.uniprot.org/] is maintained by the UniProt Consortium, formed in 2002 by the pooling of resources and expertise between the EBI, PIR (Protein Information Resource) and SIB (Swiss Institute of Bioinformatics). It is dedicated to maintaining protein sequence data and associated annotations of various sorts (functional roles, ontology annotations, sequence features, bibliographical references, links to other, more specialised databases...). It contains several modules with different objectives:

- UniProt KnowledgeBase comprises two parts, previously known as SwissProt (database of protein sequences with manually-curated functional annotations) and TrEMBL (database of protein sequences translated from gene sequences with mainly computationally predicted annotations).
- UniRef clusters protein sequences into families in order to speed up sequence similarity searches.

- UniParc is an archive of past sequences and annotations.

The latest UniProt developments are presented in [32].

An initiation to the differences between these various resources can be found at [http://www.ncbi.nlm.nih.gov/books/NBK21105/#ch1.Appendix_GenBank_RefSeq_TPA_and_UniP].

Though many bioinformatics resources also propose research-orientated sequence data banks and/or bases, the implicit requirement of referencing corresponding entries in the main resources presented here is generally respected. Actually associating functional data with genome or protein sequences is called functional annotation, as is discussed in chapter IV.

Many genome- and metagenome-sequencing projects exist today that promise riches of genomic sequence data. However, knowing genome sequences without knowing their function is of little use for understanding the dynamics of life. Genomes and genes must thus be functionally annotated to be of any use.

III.B. Annotation strategies and platforms

III.B.1. The different levels of genome annotation

Genome annotation is the process of assigning **biological roles** to nucleotide spans in the genome. Three levels of genome annotation exist, relating to three levels of complexity [33,34].

Syntactic annotation: this level deals with locating genome stretches of particular interest (typically Coding Sequences “CDS”, which are the portions of a gene's DNA that actually encode the amino acid sequence of a polypeptide or **protein**). Namely, identifying which ones actually correspond to true coding genes (a process referred to as gene calling), which ones are dedicated to transcription regulation, etc.

Functional annotation: this level assigns functions to genes of interest detected during syntactic annotation, and includes (but is not limited to) metabolic activities. As said, functions can belong to biological processes described at varying levels of detail. Three levels can be distinguished: a) **molecular function**, capturing the biochemical or structural role of the coded protein or ribozyme; b) **cellular function**, which describes the gene product's role in a higher-level cellular process, such as a metabolic pathway for an enzyme-coding gene; c) **phenotypic function** includes function of the gene product at the systemic level, and takes into account organism-wide effects stemming from gene modifications. Free text or preferably functional ontologies are suited to describing molecular functions of genes. Some relevant details of functional annotation are developed in part IV.A.

Relational annotation: this is the level that describes the relationships between all the objects and functions previously found. It is focused on building contextual representations of previous knowledge, such as inserting an enzymatic activity into a metabolic pathway (see part V.A), highlighting a gene's position in a gene expression network, or establishing evolutionary relationships across multiple gene families. Such annotation can be exploited to improve upon itself, as we shall see in chapter IV.

III.B.2. Annotation platforms: ultimate annotation tools

Before bioinformatics became a scientific field in its own right, all three levels of genome annotation had to be done sequentially by the manual and painstaking work of geneticists, biochemists and molecular biologists, exploiting available **experimental evidence**.

Today, under the pressure of the sequence data deluge, many resources and tools have been developed to ease and speed up genome annotation. The foremost of these developments is the possibility of computationally storing and representing a genome, its located genes and their associated functions. Experimental data has since been used to populate **databases** upholding various **data models**. Bioinformatics **tools** have been developed to **predict** annotations on the basis of sequence data. Gene calling, function prediction, and relational annotation can now all be carried out by automated programs to some extent. However, manual expertise continues to be required to evaluate, compare and combine the results of these predictive methods into clear, correct annotations.

Annotation platforms are collections of bioinformatics data, models, tools and interfaces made accessible to the scientific community in order to help bioanalysts create new or improved annotations by taking advantage of existing syntactic, functional and relational annotations [33].

Each platform necessarily includes three components:

- **Data management system:** large quantities of biological data, such as nucleic acid or amino acid sequences, but also gene/protein annotations, organism phenotypes, cross-references to other bioinformatics resources, pre-calculated results for heavy bioinformatics methods... obviously require proper management in order to be stored, traced, exchanged and accessed. Flat files, file formats, and databases are at the heart of this component.
- **Production system:** at the most basic computer level, a genome is little more than a file describing its sequence in letters. Actually parsing this crude data into the data management system and transforming it into something useful for scientists is the objective of the production system, which can be more or less automated. The system is generally in the form of a pipeline that streams data from the initial genome sequence to the syntactic annotation methods, the results of which the functional and relational annotation methods can build upon.
- **Visualisation system and analysis tools:** A genome that has been parsed by the production system into the data management system must be made accessible to the scientists that intend to use it. Programs or web interfaces are thus conceived to interrogate the data management system and analyse the results in a transparent way for users. Some interfaces allow bioanalyst users to enrich or improve the data contained within the underlying data management system.

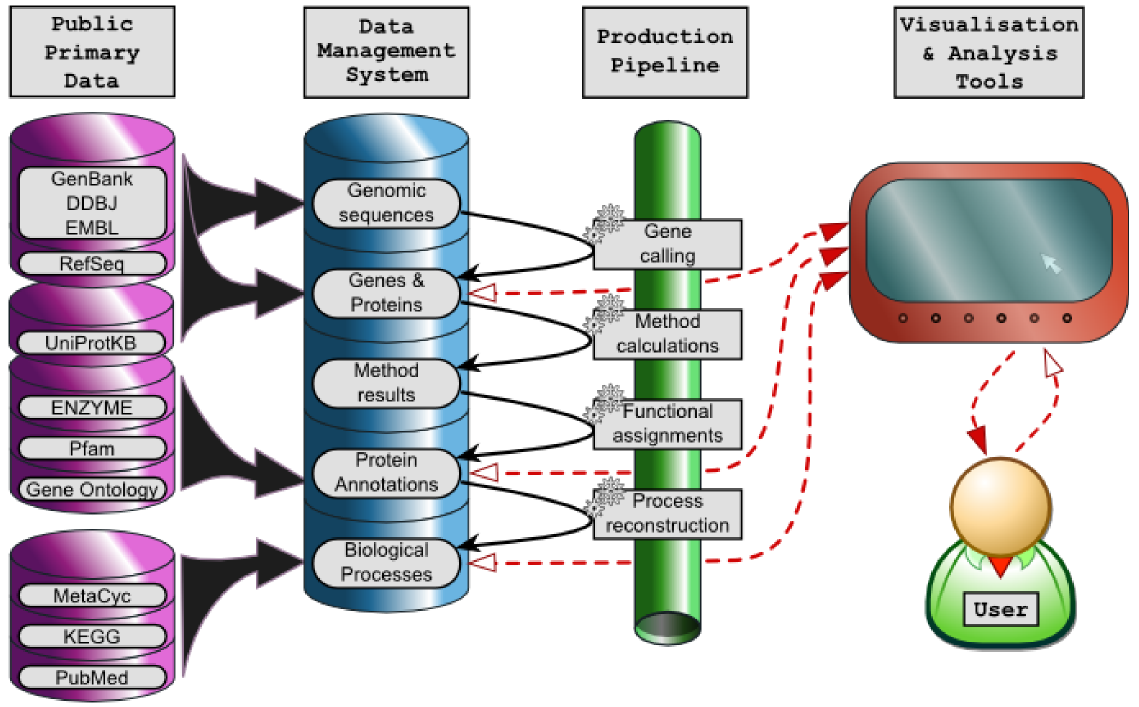


Illustration III.1: General structure of a bioinformatics platform

Publicly-accessible databases (such as the selection illustrated on the left) feed the data management system with primary data. The production pipeline sequentially transforms the data into higher-level, secondary data (from genome sequences to gene sequences to annotations to bio-processes). The visualisation and analysis tools allow platform users to interrogate the primary and secondary data (full red arrows), and possibly to generate novel data that is re-injected into the platform (empty red arrows).

Many available bioinformatics resources contain one or combine several of these components. The Generic Model Organism Database (GMOD) project [www.gmod.org] is a well-used collection of freely-available software tools encompassing data management (via flat files or relational databases using Chado [35] or BioSQL [www.biosql.org]) and visualisation/analysis tools such as the ARTEMIS viewer developed at the Sanger Institute [36]. Listing *all* such resources is not pertinent to this thesis, however I shall cite several full-blown *prokaryote* annotation platforms of note, to serve as landscape references.

III.B.3. Noteworthy prokaryote genome annotation platforms

A large number of such platforms exist and are currently used to annotate prokaryote genomes, each with their own specificities. Listed below is a selection of these:

- **Integrated Microbial Genomes** (IMG) is a comparative genomics annotation platform dedicated to annotating prokaryote genomes hosted by the United States Department of Energy [37].
- **ERGO** is a professional (and thus fee-requiring) comparative genomic platform boasting high organism and function coverage [38].
- **The SEED** is an annotation platform maintained by the Fellowship of Genomes and is centred on a gene-family-wide annotation approach, with manually-defined functions grouped into biological subsystems [39].
- **AGMIAL** is an INRA-hosted system for bacterial genome annotation which strives to offer a modular approach to annotation [40].
- **GenDB** is a bioinformatics platform maintained by the Center for Biotechnology of the Bielefeld University (Germany). It is dedicated to the automatic and expert annotation of prokaryote genomes. It is particularly rich in ontology sources (see part IIV.A for details on functional ontologies).
- The former TIGR institute hosted a prokaryote genome expert annotation platform. At the merger of the institute with others, the platform has since been split into two child platforms maintained by different teams. The Institute of Genome Sciences (Maryland, USA) hosts one of these, called the "**Annotation Engine**", to which researchers may submit prokaryote genomes in FASTA format [http://ae.igs.umaryland.edu/cgi/index.cgi]. The J. Craig Venter institute provides the "**Annotation Service**", as well as their consultative Comprehensive Microbial Resource (CMR) website. Both platforms use **MANATEE**, a freely-available web-based tool for the manual annotation and analysis of data produced by a production pipeline.
- **MicroScope** is a comparative genomics annotation platform dedicated to annotating prokaryote genomes and is hosted by the French National Sequencing Center, the Genoscope [4]. MicroScope's web-based visualisation system is called MaGe (Magnifying

Genomes) [6] and its data management system is called PkGDB (Prokaryote Genome DataBase). It contains a local installation of a collection of species-specific metabolic pathways called MicroCyc. Further details on the inner workings of MicroScope relevant to the works in this thesis are given in the following section, as well as in parts VII.D.1 and VII.F.1.

III.B.4. MicroScope Platform Overview

As for all bioinformatics platforms, there are three major components in the MicroScope platform :

- **Data management system**: the relational database system “PkGDB” (prokaryote genome database) stores all MicroScope's sequence and non-sequence data in object- or relation-specific database tables[5]. This data includes data from public databases and pre-calculated results from the production system, including MicroCyc, a pathway genome database resource initially based on MetaCyc. This component is detailed in part VII.D.1.
- **Production system**: the script pipeline is handled by a Java-based automated controller, and regularly re-runs a large set of bioinformatics methods on MicroScope genomes.
- **Visualisation system**: the web interface “MaGe” (Magnifying Genomes), which allows users to access all PkGDB data via interactive graphical representations and data tables. This component is detailed in part VII.F.1.

One of the main particularities of the MicroScope platform at its time of conception was the inclusion of pre-calculated **syntons** (*i.e.* groups of genes conserved across two genomes). MicroScope also allows users to annotate genes in a relatively controlled way and conserves annotation history for further reference, ensuring that human expertise continues to improve the data. Users are encouraged to cite articles by their PubMed IDs when appointing an annotation, ensuring that experimental evidence may be shared across annotations.

Armed with these annotation resources, tools and interfaces, bioanalysts can look to discovering the functions of prokaryote genes. Actually establishing an accurate way of defining gene function is,

[5] *Terminology remark*: the entire PkGDB “database system” contains multiple child “databases” that are each dedicated to specific data. One of these is called “pkgdb” itself, and contains the core MicroScope tables. I shall strive to make distinctions clear in this manuscript.

however, a non-trivial subject. In the following section, I shall thus briefly review what gene function is, and ways of defining it. Furthermore, a plethora of methods exist that aim to uncover the functions of genes from newly-sequenced genomes, adding to the complexity of functional annotation. I shall present some of the core concepts behind these methods, before proposing a classification of automatic annotation methods to help understand their diversity.

IV. Functional annotation

DNA sequences can be read and copied into RNA sequences by RNA polymerases and associated enzymes, a process known as **transcription**. Some of these RNAs, called mRNAs (for messenger RNAs) are then **translated** by special molecules, ribosomes, into **proteins**, using the three-base-to-one-amino-acid codon correspondence known as the genetic code. Each protein is thus created as a sequence of amino acids that then folds (either spontaneously or aided by other biomolecules) into a complex three-dimensional structure. Some proteins remain as single structures, others are assembled together into multi-protein complexes. A cell's genome thus encodes all the necessary information to continue its own life [41].

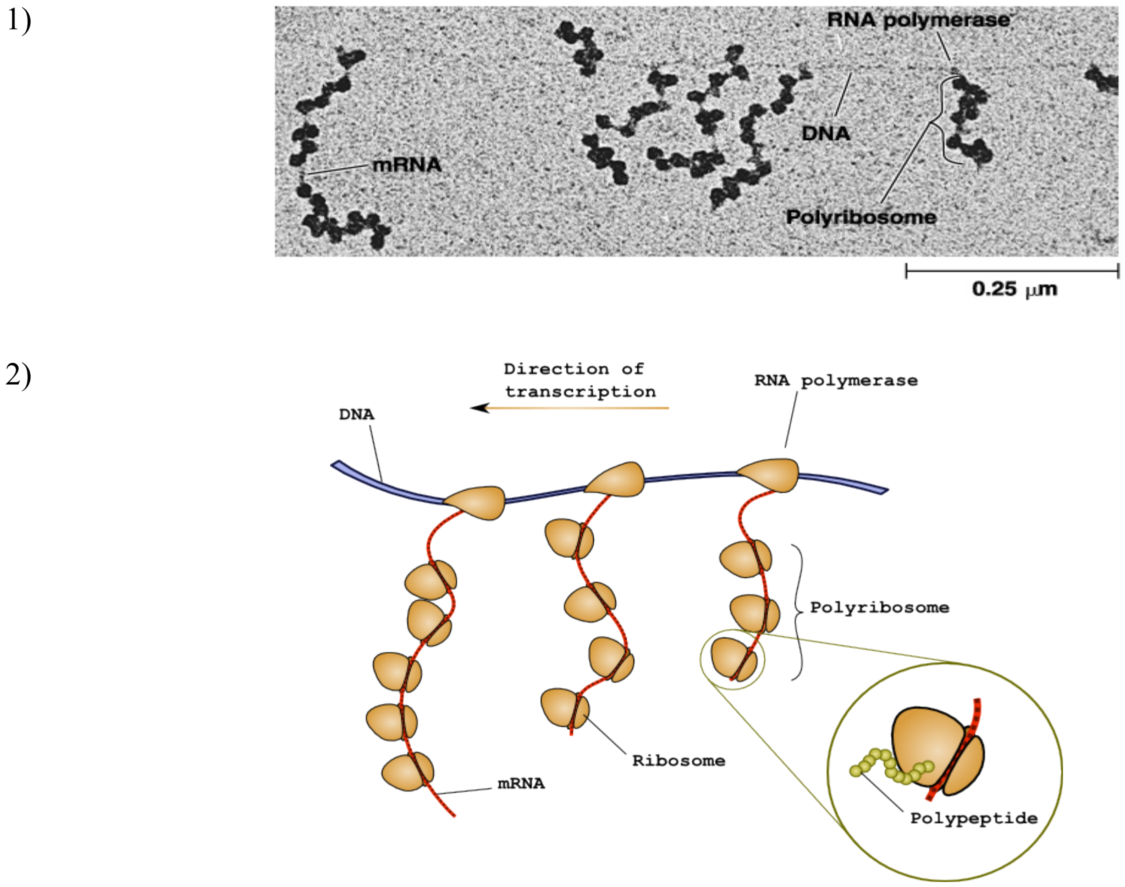


Illustration IV.1: Prokaryote transcription and translation

In prokaryotes, transcription and translation can occur at the same time. 1) Electron Micrograph of a strand of DNA being transcribed by RNA polymerases, generating messenger RNA (mRNA) that itself is translated into polypeptides (not visible) by ribosomes or polyribosomes. 2) A schematic representation of this process. Reproduced following course material by Steven M. Carr.

These synthesised proteins can fulfil many roles in a cell: structural proteins make up the matrix which holds the cell together; transporters allow the passage of specific molecules through cell membranes; signalling proteins can pass on chemical messages across the cell; chaperones help other proteins fold into correct 3D structures, enzymes work the chemical transformations of life, etc. An estimated 30% of any cell's dry matter is protein. Even non-proteic constituents of a cell require preparation or transformation by proteins. Proteins and protein functions are thus central to cellular life.

Webster's online Dictionary [http://www.merriam-webster.com/dictionary/] describes **function** as "any of a group of related actions contributing to a larger action, especially the normal and specific contribution of a bodily part to the economy of a living organism". Gene functions are all the various roles each gene product (be it RNA or protein) can fulfil *in vivo*, and at diverse abstraction levels, from specific molecular role to cell-wide high-level processes [42].

Scientists from many disciplines (molecular biology, genetics, biochemistry, medicine... to name but a few) are likely to generate functional annotations for genes and proteins during the course of their work. Different types of annotation, different professional origins and personal preferences shape the way they build their annotations, leading to wide variations in terminology that hamper information storage and retrieval from computerised systems [43]. Several efforts have been undertaken in order to develop a widely-used, precise and computationally meaningful **ontology** (*i.e.* a controlled vocabulary) to solve this problem.

IV.A. Gene/Protein function ontologies

Several relatively-used functional ontologies are based on one of the first efforts, presented in [44].

MultiFun: The first ontology based on the initial works of Riley *et al.* is **MultiFun** [45], an ontology designed primarily for use in *Escherichia coli*. Main gene product function categories include Metabolism, Information Transfer, Regulation, Transport, and Cell Structure. The current MultiFun classification is available from [http://genprotec.mbl.edu/files/MultiFun.html].

MIPS FunCat ontology: this ontology is described as a hierarchically structured, organism-independent protein function classification scheme applicable to all domains of life, though it was designed specifically with Yeast in mind [46]. In order to retain a minimal level of descriptive quality, FunCat specifies protein activities only down to a broad functional level (*e.g.* biosynthesis of glutamine), rather than a specific role (such as glutamine synthetase). FunCat terms are thus more process- and pathway-related than activity-related.

JCVI ontologies: The J. Craig Venter institute (JCVI) uses two types of ontologies. The **JCVI roles** (previously TIGR roles) are a two-level functional classification system of protein cellular functions based on Riley's works [47], assigned on the basis of proteins belonging to JCVI protein families (TIGRFAMs). The **JCVI Genome Properties** are organism-level hierarchical descriptions of taxonomic, phenotypic, and predicted traits [48].

UniProt: The protein sequence database uses a **keyword**-based controlled vocabulary for rapid annotation and retrieval of the functional data associated to its protein sequences. The keyword categories are Biological process, Cellular component, Coding sequence diversity, Developmental stage, Disease, Domain, Ligand, Molecular function, Post-translation modification and Technical term [49].

Gene Ontology: One of the well-used systems for describing gene functions (and not limited to metabolic functions) can be viewed as a generalisation of all previous attempts. **Gene Ontology (GO) terms** [43] are terms that are organised hierarchically, each level providing additional functional detail. However, unlike FunCat, multiple inheritance and multiple types of term relationships are allowed, giving this scheme much more flexibility, at the cost of increased complexity. Belonging to a structured ontology, GO terms are much more informative and easier to handle computationally than **Enzyme Commission (EC) numbers** (for details on this strictly metabolic annotation terminology, see section V.A.5.b) for example. They are organised into three main directed acyclic graphs: molecular function, biological process, and cellular localisation. The current trend in bioinformatics is to annotate genes with GO terms, and many automated prediction methods now output GO terms. An example set of GO terms is represented below using Cytoscape [www.cytoscape.org]:

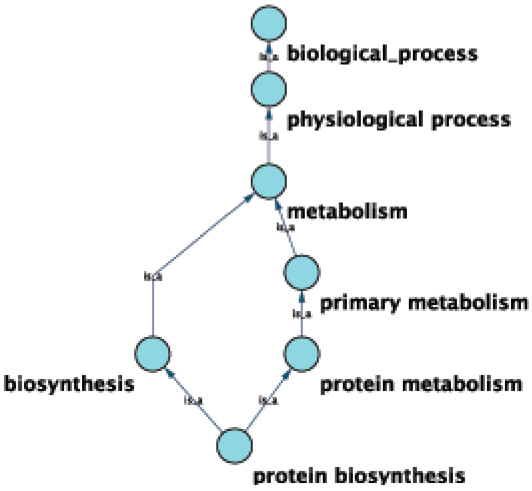


Illustration IV.2: An example of the hierarchical graph structure linking Gene Ontology terms.

Each Gene Ontology term can inherit from one or several other terms. Here, protein biosynthesis is a descendant of the head term “biological_process” via two different paths.

Ontologies like GO terms describe gene and protein functions in various levels of detail, but are not adapted to linking them to the actual chemical transformations carried out by enzymes, and hence are of limited use when studying an organism's metabolism *in fine*. Detail on metabolism-specific alternatives are given in part V.A.5.

Having defined the terminology with which genes can be annotated, it remains to be seen how annotations are actually assigned to a given gene. Many annotation protocols actually require retracing the evolutionary history of several proteins. I shall thus detail how protein evolution can be of importance to functional annotation, before discussing a classification of the different types of annotation that exist.

IV.B. Key concepts in functional annotation

Many mechanisms are known today that can account for the accumulation of punctual or large mutations in genetic material over the course of many generations, and most are of common knowledge. Analysing the sequences of several genes, in the light of these mechanisms, can shed light upon the evolutionary relationships between them, as described in the following section.

IV.B.1. Phylogenetic links

Homology, orthology, and paralogy: Historically, homology was used by comparative biologists to describe an evolutive link between organs or traits of different species. Structural and functional similarities between organs allowed scientists to uncover these links. Two organs from two species thought to originate from an ancestral organ in a common ancestor to these species were said to be homologous (*e.g.* flipper of seal and hand of human). Similar organs or traits that evolved separately to fulfil the same function are referred to as analogous (*e.g.* wing of bird *versus* wing of bat).

These notions have been transferred to genetics. Two genes (or gene products) of different organisms are said to be **homologous** (*i.e.* are homologs/homologues - for details concerning spelling see references towards the end of this section) when sufficient evidence has been found to assert that both genes have evolved from a common ancestral gene. **Analogous** genes have similar molecular functions but have evolved separately. The typical working hypothesis is that homologous genes encode proteins with identical or at least related functions. Put simply, two genes having diverged over evolution by only a few base pairs lead to proteins diverging in only a few amino acids, and thus to functions with very little difference. Obviously, the non-linear links between gene sequence, protein sequence, protein 3D structure and activity cloud these relationships somewhat.

Different types of gene homology can be observed, corresponding to different evolutionary paths which led to different selective pressures being applied to the genes. These different pressures limit the possibilities for the evolution of these genes. Thus, actually inferring the conservation of low- or high-level gene function from homologous genes requires identifying the specific type of homology involved.

A **speciation event** is a complex event (usually including many mutations accumulated over several generations) that leads to the creation of two new species from a single ancestor specie. Due to the common ancestry, most of the genes of both species have common ancestral genes in the ancestor specie. Genes having a common ancestral gene with which they are separated only by speciation events are known as **orthologs** (orthologues). Orthologous genes are supposed to have been submitted to the same selective pressures in both lineages, ensuring conservation of the genes' functions.

A **gene duplication** event leads to the creation of two copies of the same gene in the genome of a given specie, that can evolve with lightened selective pressures. These genes are known as **paralogs** (paralogues), and the lesser selective pressure is not considered to conserve gene function, though paralog functions can be similar (*e.g.* enzymes with different substrate specificities, see part V.C.1).

Both gene duplications and speciation events can line the evolutionary paths of a pair of genes, making it difficult to classify them as orthologues or paralogues. Different cases arise depending on whether a given speciation took place before duplication, or after. A duplication pre-dating a species' split is considered “ancient” enough so that the genes it gave rise to have diverged in function; these genes are called **out-paralogs** (a terminology inspired from phylogenetics). Genes born from a post-speciation duplication are considered recent enough to have not diverged functionally; they are called **in-paralogs**, and can be considered as orthologs (they are sometimes referred to as co-orthologs) [50]. When considering more than two genomes (and thus more than one speciation event), in- and out-paralogy becomes dependant on the chosen reference speciation [51]. In all cases, the number of generations since the duplication event should be considered, in order to evaluate how “recent” or “ancient” it really is. These concepts are illustrated in the figure on the following page [Illustration IV.3: Homology, orthology, paralogy over three related species].

Further complicating evolutive history is the possibility of Horizontal Gene Transfer (HGT). This phenomenon occurs when a stretch of DNA (containing potentially one or several genes) is transferred from an organism from one specie to one of another specie. HGT mostly concerns prokaryotes and single-cell eukaryotes, though some evidence has been reported showing HGT concerning higher eukaryotes, *e.g.* [52]. Homologs suspected or proven to have been acquired by HGT during evolution are referred to as **xenologs**. Xenologous gene displacement occurs when a xenolog gene is acquired in a genome and the original gene is lost, effectively replaced.

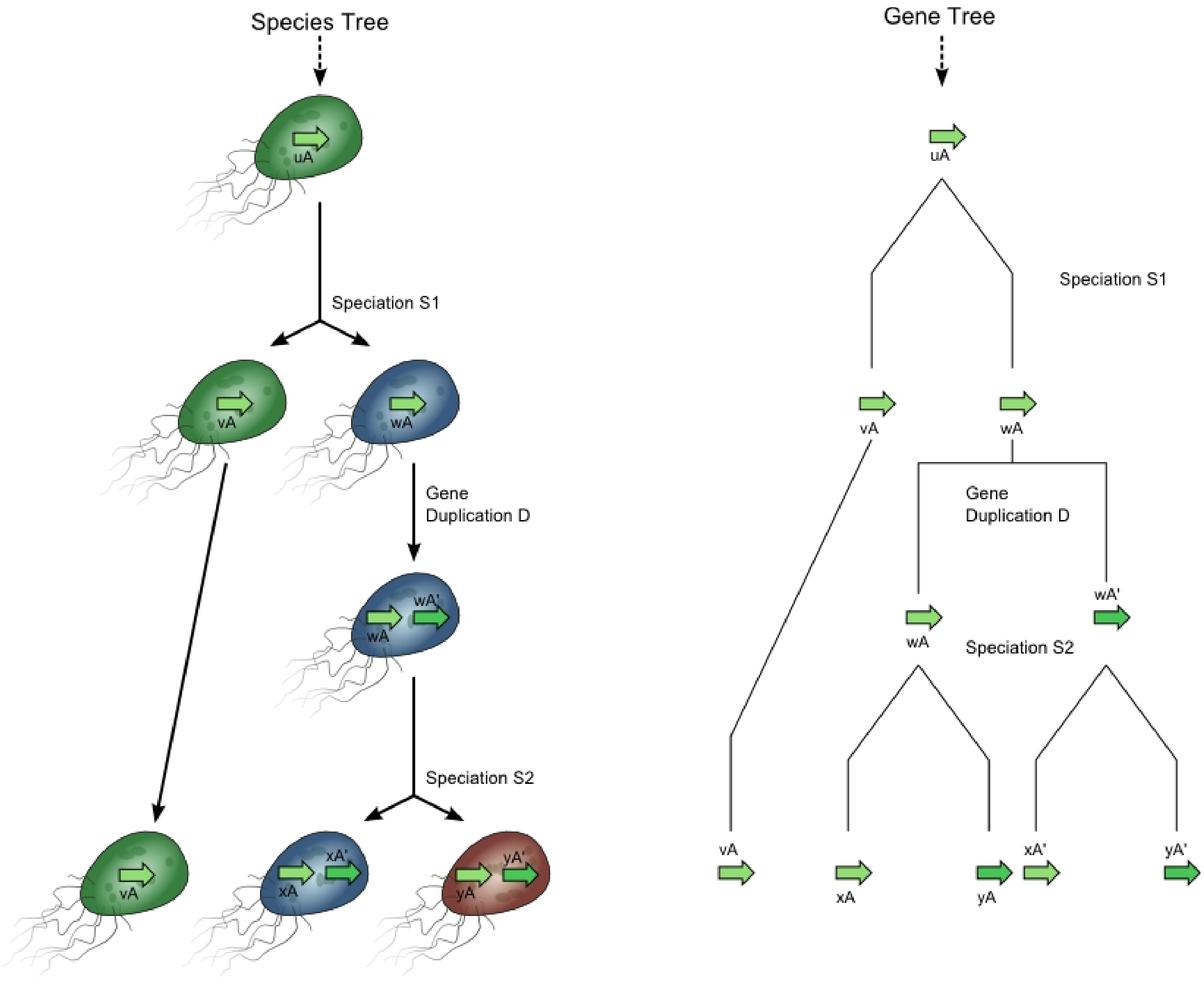


Illustration IV.3: Homology, orthology, paralogy over three related species

Homology: The genes vA, xA, xA', yA, yA' are all homologs as their evolutionary history can be traced back to a common ancestor gene (uA).

In- and out-paralogy: Gene pairs (xA, xA'), (yA, yA') are reciprocally paralogs (as well as the ancestral (wA, wA')). Actually specifying in/out-paralogy depends on the chosen speciation event. For instance, genes xA and yA' are in-paralogs with respect to speciation S1 (duplication post-speciation), but are out-paralogs with respect to speciation S2 (duplication pre-speciation).

Group of orthologs: In the case illustrated here, it is not possible to partition the genes into groups of orthologs and in-paralogs with respect to the last speciation event (S2). Indeed, vA is orthologous to all other genes, but they do not form a group because every other pair is out-paralogous with respect to speciation S2.

Many of the terms defined here were first launched in genetics in the seminal paper [54]. They have further been discussed, along with their spelling (*e.g.* homologue *versus* homolog), in many papers and comments, the most amusing of which may be the series [55–57]. [52] also provides (amongst other things) an interesting review of the question.

IV.B.2. Conserved sequences, domains, and gene/protein families

One type of hint to the homology between genes or gene products is that of **conserved sequence**. Indeed, given that nucleotide or amino acid sequences evolve over time, similar sequences can be thought to belong to genes or gene products of common origin. Sequence similarity is often analysed in order to detect homologs across two or more organisms, and is the historical approach at the basis of bioinformatics today (more details in section IV.C.2). It is, however, prone to over-interpretation, leading to false homology detections [53]. Others have also pointed out that pure sequence-based approaches are reaching an informational threshold [1].

A slightly more function-orientated homology signal is that of conserved **protein domains**. The modular nature of protein structures became experimentally apparent in the early 1970s. X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy revealed that proteins contained structurally distinct parts that were later shown to be units of protein folding, function, and evolution. The 3D elucidation of protein structure being a relatively long and complicated experiment to perform, bioinformatics were called in to help. The alignment of multiple sequences containing similar domains showed that the amino acid sequence of the domains was generally much more conserved than that of the rest of the proteins. The definition of protein domains has grown to include stretches of conserved sequences (even non-contiguous stretches) that may or may not have a basis in particular 3D structures. Today, the crossover between manually-curated experimental 3D structure-defined domains and automated sequence alignment-defined domains is such that some care must be taken in distinguishing the origin(s).

Conserved domains, given their relationship to gene product function, can be a strong marker of evolutionary relationship that conserve function, hence their use in determining homologous **gene/protein families**.

A loose way of finding groups of genes/proteins with homology relationships is establishing gene/protein families. Gene families can be defined in two ways. The first is based on identifying domains within the sequences or structures; proteins with similar domains or domain architectures can be expected to be related functionally, and form **domain-based protein/gene families**. The other requires establishing evolutionary relationships between proteins (or their coding genes) in order to infer shared function. Once these links have been untangled, it is often of interest to regroup genes into clusters sharing a given history and hopefully being relatively iso-functional (*i.e.* sharing a same function). These clusters are called **(sequence-based) protein/gene families** and are generally based on orthologs, as well as added in-paralogs ("recent" paralogs) that are considered to not have had the time to diverge significantly, while taking care to exclude out-paralogs ("ancient" paralogs) that have probably diverged.

If domains are units of folding, evolution, and function, then the entire 3D structure of a protein can also be relevant to these notions. As such, comparing 3D structures can shed light on otherwise-invisible evolutionary relationships (or their absence). It is thus possible to generate protein families based on 3D structure conservation, rather than on sequence feature conservation.

Families are not the only way of exploiting evolutionary clues. A less direct way involves the use of contextual information across many different genes in order to capture higher-level evolutionary pressures, as presented below.

IV.B.3. Functional dependence

Genes whose products participate in a same biological process are likely to be submitted to common evolutive, regulatory and physical constraints. Indeed, if a complete process contributes to the fitness of an organism (an evaluation of how adapted it is to its environment and how likely it is to produce offspring), then loss of one or more parts of the process are as detrimental as the loss of the whole process. Also, the entire process can evolve, rather than just parts of it. Such an evolutive pressure can be identified using several sources, such as gene clustering, syntenies, phylogenetic profiles and gene co-expression (definitions and detection methods will be given in part IV.D.2 and later). Functional dependency is also referred to as "guilt-by-association" [54].

As all these methods rely on identifying clues from information *between* or *across* genes (*i.e.* contextual information), rather than within the target gene to annotate. They are thus referred to as **context-based methods**. They are opposed to sequence similarity-based methods, though obviously the knowledge of the sequences of all concerned genes is, at some step or other, required. For this reason, I will always consider that the full genome sequence is available when describing these methods in the rest of this manuscript.

IV.B.4. The three types of gene functional annotation

We have seen that, by taking evolutive history into account, the functions of some genes can be clues for the functions of others. I propose that annotation methods be classified into three main types: **experimental validation**, **homology-based transfer**, and **functional dependency-based inference**. Experimental validations can demonstrate the function of a gene with high confidence, and are precious starting points for methods of the two other types. As suggested in [55], these are bioinformatics methods: one type is based on the transfer of existing annotations between genes of different organisms on the basis of detected homology; the other infers possible annotations from the annotations of other genes of the same organism on the basis of detected functional dependency.

An annotation is experimentally validated when a scientific **wet-lab experiment** (be it *in vivo* or *in vitro*) demonstrates the function of the gene or genes in the studied organism. Many different protocols exist that can help build these demonstrations, such as **gene invalidation**, **mutagenesis, atom tracing**, **expression cloning**... This kind of result is usually published in an **article** which serves as a reference for the annotation. For example, the MicroScope platform specifically allows bioanalysts to flag genes with citations of papers that form the basis of its annotation.

The three types of annotation ;are illustrated on the following page Illustration IV.4: The three main types of functional annotation.

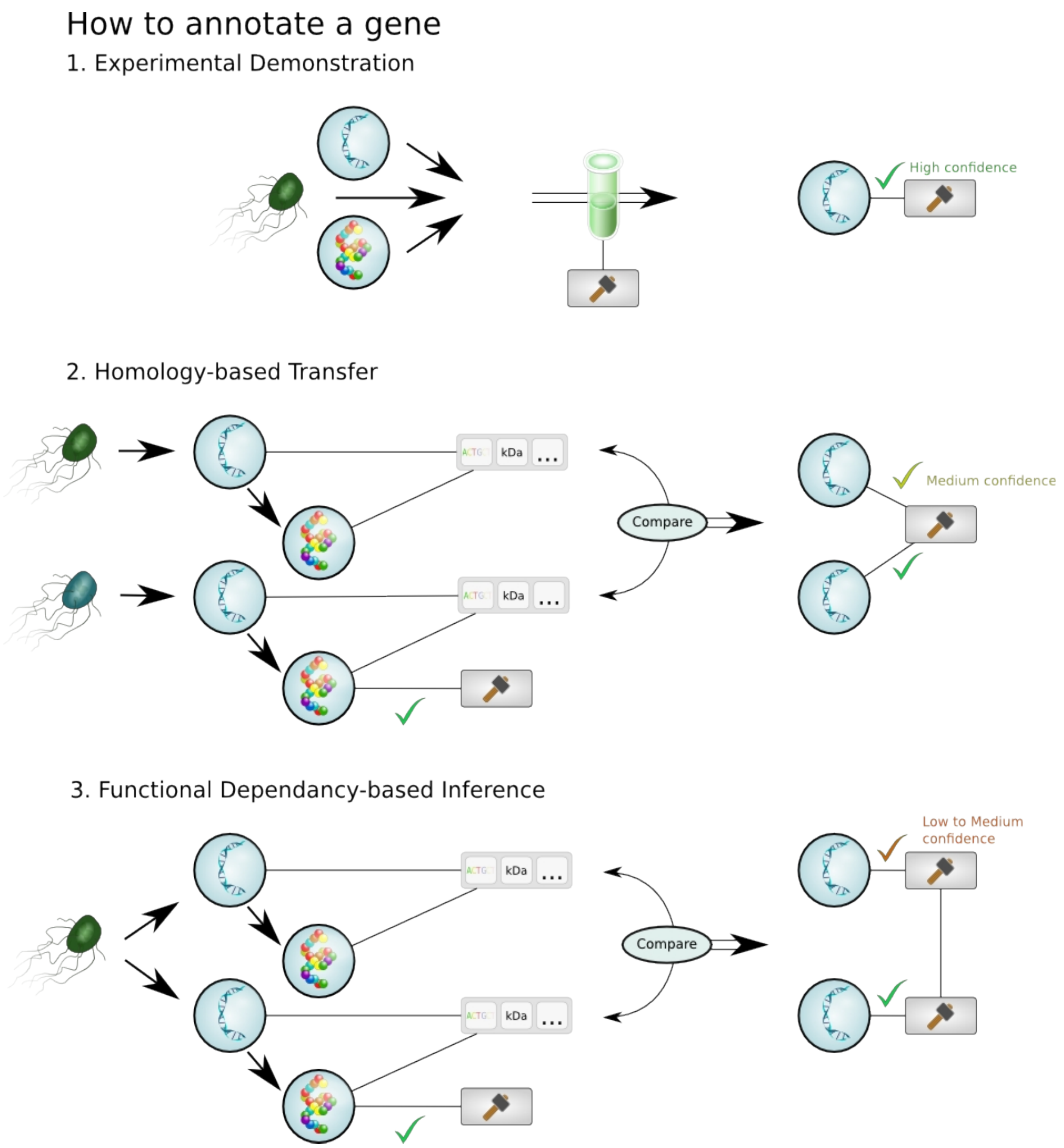

How to annotate a gene
1. Experimental Demonstration
High confidence
2. Homology-based Transfer
ACTG
kDa
...
Compare
Medium confidence
3. Functional Dependancy-based Inference
Low to Medium confidence

Illustration IV.4: The three main types of functional annotation

Experimental validation (1) allows the high-confidence annotation of a gene with a given function. Lower confidence annotations can be generated by homology-based transfer (2) and functional dependency-based inference (3), which both rely on the comparison of gene or protein features (sequence, physico-chemical properties, genomic context...). Said features concern genes in the same or in different organisms for the former, and genes of the same organism for the latter.

As already said, the experimental association of a gene/protein sequence with a specific activity is tedious and costly; hence the utility of a) transferring previous annotations to new genes on the basis of detected homology, or b) using them in the context of inference. I shall now present these two approaches.

IV.C. Annotation transfer on the basis of homology

IV.C.1. Protocol

As described, the functions of homologous genes are thought to be near-identical (for orthologs), similar (for recent paralogs), or vaguely similar (for ancient paralogs). Gene homology can be inferred upon analysis of gene data.

The field of bioinformatics was born with the development of a method for establishing gene homology on the basis of gene or protein sequence. Indeed, genes evolve through the accumulation of mutations in their sequence, so comparing gene sequences should allow one to reconstruct gene evolutionary relationships as a phylogenetic tree. Identifying paralog cases requires comparing multiple sequences from various species. To avoid problems posed by the redundancy of the genetic code, one can also directly use amino acid sequences, which additionally are more closely related to the protein's function.

Once homology has been established between a pair of genes, this new knowledge can provide support for the functional annotation of one gene, using previous information about the function of the other. This is referred to as "**annotation transfer**".

With the exponentially increasing quantity of sequence data that is available to the scientific community, traditional experimental annotation of gene function for all genomes has become infeasible, spurring on the development of bioinformatics methods for gene annotation. Many sequence-based approaches have been imagined in order to establish homology relationships to help annotation transfer. Those based on phylogenetic tree construction and analysis are generally regarded as too computationally intensive, and often require manual expertise, making them impossible to use on a large-scale basis. As already described, using simplifications or proxies has thus become standard. Some methods are based on establishing gene/protein families by comparing the sequences of all known genes/proteins and clustering the most similar; others look for conserved amino-acid domains that can be linked to elements of protein function; some use a more information-theoretical approach and count amino-acid motifs... Even predicting protein physico-chemical properties, sub-cellular localization, secondary or tertiary structures from sequence gives insight into how two gene (products) can be related. In all cases, it boils down to analysing the similarity between sets of sequence-derived features describing two or more genes [56]. In essence, "sequence-similarity based methods" are all homology-determining methods.

IV.C.2. Homology-determining methods

In bioinformatics, the historically-used evidence of homology is that of sequence (nucleotide or protein) similarity (and they are often used synonymously, which is a malapropism), which can be established by using sequence alignment programs such as **BLAST** [57], FASTA [58] or the Smith-Waterman algorithm [59]. Slightly stronger than a good alignment hit is the notion of bidirectional best hit (BBH)[6], wherein each gene scores its highest alignment score with the other, all other alignments with genes from both genomes considered. More recent developments go beyond simple sequence alignment, including PSI-BLAST [60] and HMM-based methods, *e.g.* [61–63].

In practice, a minimum BLAST-established sequence identity of 35-40% over a good match length is required between two prokaryote protein sequences for homology to be inferred by this method, though in the presence of many other clues supporting identical function, this number has been known to go as low as 10-15%. Applications of sequence similarity-based annotation transfer to whole genomes has helped the annotation of little-studied organisms, *e.g.* in [64,65], as well as speeding up the annotation of all newly sequenced organisms..

Functions can also be transferred between members of hopefully iso-functional families, or to newly added members to the family. David VALLENET and I distinguish three approaches to building families, a classification detailed in this Chapter: *de novo* sequence clustering, semi-supervised sequence similarity-based aggregating, and semi-supervised domain-based aggregating.

Here, rather than present all available procedures than can be used to build gene/protein families and transfer annotations on that basis (which is beyond the scope of this thesis), I shall present a select few that are linked to the works in this thesis, either by their intervention in the CanOE strategy (chapter VII), the BKACE project (chapter VIII), or simply because of their use in the MicroScope platform.

IV.C.2.a. De novo sequence clustering

These methods rely on sequence similarity alone to cluster genes or proteins together into new families. The addition of novel sequences would ideally require recalculating all families, though sub-optimal incremental addition procedures can be used.

[6] a.k.a. reciprocal best hits, or variants thereof.

InParaNOId: The InParaNOId algorithm [50] was designed to detect ortholog and co-ortholog proteins between two eukaryote genomes, more specifically with the then nearly-completed human genome in view. The algorithm uses all-*versus*-all protein BLAST results (only keeping alignments whose bit-score and overlap are over a minimum threshold) to detect mutually best hits that are considered as seed ortholog pairs. One out-group genome can be included in order to refine this detection. Seed ortholog pairs are then augmented with co-orthologs (in-paralogs) by including same-genome proteins whose similarity to one of the proteins from the seed pair is at least as great as the similarity between the pair. These similarity scores are used to derive “confidence scores” for co-ortholog addition. Overlapping protein assignments to seed pairs are resolved using several tailored rules, leading to the definition of high-confidence gene/protein families. Functions can then be reliably transferred across members of a family.

InParaNOId was the first algorithm, to my knowledge, that specifically distinguishes orthologs and various types of paralogs, in order to improve family construction.

Tribe-MCL: This algorithm uses all-*versus*-all protein BLAST results from multiple organisms [66]. Unlike InParaNOId, any number of genomes can be used. The -log10 of the e-values are parsed into a similarity matrix which is symmetrised. This matrix is then passed to the Markov Clustering algorithm (MCL) developed by Stijn van Dongen [67]. The algorithm then quickly and efficiently partitions[7] the proteins into clusters, *i.e.* gene/protein families.

Tribe-MCL is a precursor to Ortho-MCL, described below.

OrthoMCL: The OrthoMCL algorithm [68] also uses BLAST results between proteins of different genomes and the MCL algorithm. As with Tribe-MCL, any number of genomes can be used. Ortholog/paralog relationships are inferred between proteins from a pair of genomes, similar in idea to InParaNOId, though with many more out-group genomes. A normalisation procedure is included to account approximately for phylogenetic proximity between genome pairs. The resulting data is exploited to build a similarity graph that is then passed to the MCL algorithm. Additional practical details for the (Ortho)MCL algorithm are given in section VII.C.2, as a variant of it is used

[7] Technically, it is possible for the algorithm to return clusters that overlap, thus not leading to a strict partition; however this is extremely rare in practice, and the algorithm wrapper removes any detected overlap [van Dongen, http://micans.org/mcl/man/mclfaq.html#olapintro].

as part of the CanOE strategy. OrthoMCL or a variant of it is also used in the IMG platform to define IMG ortholog groups [69].

CORRIE: the Correspondence Indicator Estimation procedure is a bioinformatics method which defines functional classes (in this case, EC numbers, see section V.A.5.b) to which it assigns proteins with known annotations, effectively generating *a priori* iso-functional protein families. For new, un-annotated proteins (or proteins to re-annotate), correspondence indicators (CIs) measuring the strength of the association of each protein with each functional class are calculated, based on the sequence similarity between the protein and protein members of the functional class. Finally, different Bayesian approaches are proposed to select, using the protein's CIs, which functional class it is most likely belong to. This procedure is described in [70,71].

IV.C.2.b. Semi-supervised sequence similarity-based aggregating

These methods rely on manual curation at some step or other in order to validate the gene/protein families that are established by sequence similarity, multiple sequence alignments and other data. Adding novel sequences to a family is relatively easy.

COG: The Clusters of Orthologuous Groups of proteins (COG) database [72] is an early and still widely-used effort to classify proteins from bacterial, archaeal and eukaryote genomes into hopefully iso-functional groups of common ascendency. Using all-*versus*-all protein BLAST results (masking certain sequence features considered to be of no interest for ortholog identification), it identifies paralogs and “collapses” them into a single representative sequence, before identifying triangles of pairwise best hits between proteins of different genomes. Such triangles are iteratively merged along shared sides into COGs. COGs are then manually analysed to reveal spurious classifications due to multi-domain proteins or other evolutionary phenomena. The final COGs are thus generated by a semi-automatic procedure and benefit from an expert curation step. No cut-offs are applied to the BLAST results, thus allowing COGs to be based on high and low similarity triplets, though they must be consistent amongst themselves. This allows COGs to capture both fast- and slow-evolving families. Adding novel sequences to a COG family is relatively trivial.

HAMAP: The HAMAP (High-quality Automated and Manual Annotation of microbial

Proteomes) project was created to allow partially-automatic procedures to help propagate Swiss-Prot (now UniProt-SwissProt) annotations with high confidence to members of a same gene/protein family, in the hope of keeping up with the continuous flow of new sequence data that expert manual analysis could not deal with [73,74]. Rather than using sequence similarity to aggregate genes or proteins, families were defined on a case-by-case basis by manual expertise. Already well-known (sub-)families were researched and used to create family-specific sets of rules that could be used to transfer annotations from actual family members to new members. These rules are based on multiple sequence alignments that are used to isolate 1) complete protein sequence profiles or meta-motifs, 2) localised sequence motifs or patterns, or 3) specific sites. These rules can combine all of these to ensure highly specific detection of proteins that belong to the same family. In this regard, HAMAP could also be classified as a semi-supervised domain-based aggregating method, and adding new sequences is simple.

FIGfam: FIGfam is another collection of protein families that are maintained by the Fellowship for the Interpretation of Genomes (FIG) [75]. FIGfam resembles HAMAP in that families are built upon expert manual curation, though functional annotations are rooted in the SEED system and thus extend beyond mere sequence similarity. FIGfam proposes an automated procedure for determining the membership of newly added proteins. Family membership is manually evaluated from the SEED [39] data using sequence similarity and genomic context.

IV.C.2.c. Semi-supervised domain-based aggregating

These methods rely on initial manually-defined families and/or manual curation of located conserved domains. All proteins containing a given domain (or set of domains) is assigned to a family, making the addition of novel sequences relatively easy.

Pfam: The Protein Family (Pfam) database is based on tracking the presence of various conserved domains within protein sequences [76]. The presence of a given domain, or a set of domains (known as a “domain architecture”) is used as a basis for defining protein families. Domains are located on the basis of high-confidence “seed” multiple sequence alignments, which are used to build a HMM profile representing the domain, which is in turn used to agglomerate additional, more distantly related proteins into what has become a protein family. Today, two types of Pfam families exist: Pfam-A families are manually-curated, whereas Pfam-B families are automatically generated and have not yet been validated. Domains that have no currently assigned function are referred to as "Domains of Unknown Function" (DUFs) and represented roughly 3,000 (22%) of all Pfam families in 2010 [77]. Pfam families are widely used and were exploited in this thesis during the BKACE project (Chapter VIII).

PRIAM: This approach is dedicated to identifying enzyme-coding genes and their enzymatic activities by using rules combining enzyme position-specific "profiles" built from collections of known enzyme-coding genes [78]. These profiles are analogous to protein domains, though positions are not necessarily contiguous. PRIAM thus assigns functions to new genes on the basis of a domain-detected homology.

InterPro: InterPro [79,80] was designed as an integrative repository for several protein-family-defining efforts (see table below). It includes documentation on families, domains and functional sites as detected by resources such as: PROSITE [81], PRINTS [82], Pfam [76,77], Blocks [83], ProDom [84], Gene3D [85], SMART[86], SUPERFAMILY [87], PIRSF [88], TIGRFAM [89], and PANTHER [90].

MicroScope paltform: all of the gene/protein family resources discussed above have been integrated into the MicroScope platform, either directly, or via InterPro.

IV.C.2.d. 3D structure-based methods

Another bioinformatics development is the use of 3D protein modelling in order to derive protein structural similarities or to carry out substrate docking experiments. These methods deal much more directly with the elements responsible for protein function (*i.e.* 3D distribution of amino acids with various physico-chemical properties) and thus hold great promise for future annotation, though obviously they are still dependant on sequence for structure elucidation. Unless otherwise stated, these approaches define protein families on the basis of 3D structure, and these families can be used for functional transfer.

The Protein Data Bank Resource: Before discussing 3D-based methods, it is necessary to cite the Protein Data Bank. The PDB is a freely-accessible archive of large biomolecule (including proteins and nucleic acids) 3D structures, maintained by the Research Collaboratory for Structural Bioinformatics (RSCB) [www.pdb.org] [91]. PDB data is available via web interfaces and can also be downloaded in the PDB file format, which describes 3D coordinates of the atoms in the structure. Such data is obviously essential to 3D structure-based homology methods.

SCOP: the Structural Classification of Proteins classifies proteins into folds, superfamilies, and families. Families and super-families are based on common evolutionary origin as revealed by sequence similarity and structures, respectively. Folds are built on conserved secondary structures, and can be additionally regrouped into classes.

CATH: CATH [92,93] is a manually-curated database of hierarchically-classified protein domain 3D structures. Structures are classified into Class, Architecture, Topology and Homologuous family. In order to build the classification, proteins having good 3D models in the PDB are grouped according to sequence similarity, domains are isolated along their sequences before being aligned in the group in order to establish protein architecture. CATH is more focused on sequence & structure similarity than SCOP. The Gene3D resource [94] is based on CATH domains.

ASMC: The Active Sites Modeling and Clustering (ASMC) method [95], developed at the Genoscope in the LABIS team, is capable of grouping modelled proteins from a given family into (hopefully) iso-functional sub-families by analysing key amino acids in their predicted active site pockets (see section VIII.C.5 for more details).

Others: [96] propose an integrated strategy, exploiting many bioinformatics resources and tools, in order to derive functional predictions from sequence and 3D structure.

IV.C.3. Limits

There are several limits to homology-based annotation transfer methods.

For a start, despite the increasing number of new genes and genomes that are sequenced, an increasingly small fraction of these are considered as representatives of novel “islands” of sequence space [1]. In this respect, homology-based methods will hit an asymptotic threshold of sequence diversity from which to transfer functions, though much work remains before the known sequence space is correctly annotated, which is in itself another limit.

Beyond simple divergence as suggested by paralogy, homology does not necessarily imply iso-functionality, given that protein functions can be multiple and diverge very quickly with few mutations. This is a painful thorn in the homology-based annotation transfer approaches, a fact that has been demonstrated and underlined in many works, such as [97–100]. Indeed, for some years now, scientists have had a hard time debunking the overly-simplistic view of “one gene - one protein - one function”. However, evidence has accumulated backing the fact that many proteins can actually fulfil not one, but many different roles in the cell. Some enzymes can catalyse sets of similar metabolic reactions, or even several completely different reactions; signalling proteins have been known to have structural roles; etc. This phenomenon is called **functional promiscuity** (see also section V.C.1). Many types of functional promiscuity exist, and several classifications thereof have been proposed [101–106]. Common to all, however, is the notion that some protein functions have been selected by evolution as the “primary” function of the protein, whereas the more numerous adventitious “secondary” functions have sprung up by chance, either within the same active site as the primary function, or in other sites anywhere on the protein. This reservoir of secondary functions is thought to be part of the great engine of evolution, allowing new functions to be tested without disrupting the cells' inner workings, and serving as a functional recruitment base when needed [107,108]. The phenomenon leading to this robustness is known as **canalisation** [109,110].

Homology-based annotation transfer methods are also subject to phylogenetic bias, as compared feature sets are most likely to be close when the host organisms are close themselves. Normalisation procedures can limit this effect, such as the protocol presented in OrthoMCL [68].

Finally, the plethora of methods for inferring homology and transferring annotations on that basis (of which only a small selection of methods have been presented here) can deter any newcomer to the field. To pair with this diversity, the evaluation of method performances is still difficult to perform and seldom done, not making choosing a given method any easier. To my knowledge, the best attempt of comparing multiple family-building methods can be found in the excellent paper [51].

Despite the best results obtained using ever-finer sequence-derived homology detection approaches, such methods obviously cannot help when searching for candidate genes for orphan reactions, as no known gene sequences correspond to the target activity. They cannot help either when an entire protein family yields no existing functional annotations as no annotations can be transferred to it. Hence the requirement for so-called “sequence-independent” context-based methods.

IV.D. Annotation inference on the basis of functional dependence

Functional inference is carried out between genes of a same target organism, and has only become feasible with the advent of whole genome sequencing [111].

Three steps are required before inferring functions from functional dependence: 1) dependence detection for all genes, 2) gene context extraction for a target gene, and 3) functional context construction for the target gene. This functional context can then be used to 4) infer possible functions for the target gene [112,113].

IV.D.1. Steps

Dependence detection: Functional inference is based on the precept that genes working together in a same biological process are submitted to common evolutive, regulatory and physical constraints. Measures of functional dependency genes attempt to capture these constraints, and their known evidence sources are listed in the following section.

Gene context extraction: Functional dependency between genes can usually be represented as weighted networks, wherein genes are nodes and functional dependency is captured in edge weights (binary or real). The neighbourhood of a target gene is a set of genes that are "close" to it in the network (*e.g.* at most k edges away). The actual context used in the next step is defined in some methods as simple neighbourhoods, in others as gene clusters based on network connectivity.

Functional context construction: Functional context can broadly be defined as the set of all functional data available for all genes from the gene context of a gene of interest. While gene context allows linking functionally dependent genes together, functional context reconstructs an image of the way the involved genes all participate in a same biological process. When annotating a gene of interest, bioanalysts manually reconstruct the surrounding gene and functional contexts in order to guess the target gene's possible activity.

Function inference: Computationally, the function of a target gene (or a set of more or less likely functions) can be inferred from its functional context either directly (*e.g.* using the 'majority rule', transferring the highest-level most common annotation in the context to the target) or indirectly (*e.g.* finding reaction gaps, using machine learning predictors, using functional similarity measures...). In the second case, it may be necessary to determine beforehand which possible functions can be inferred.

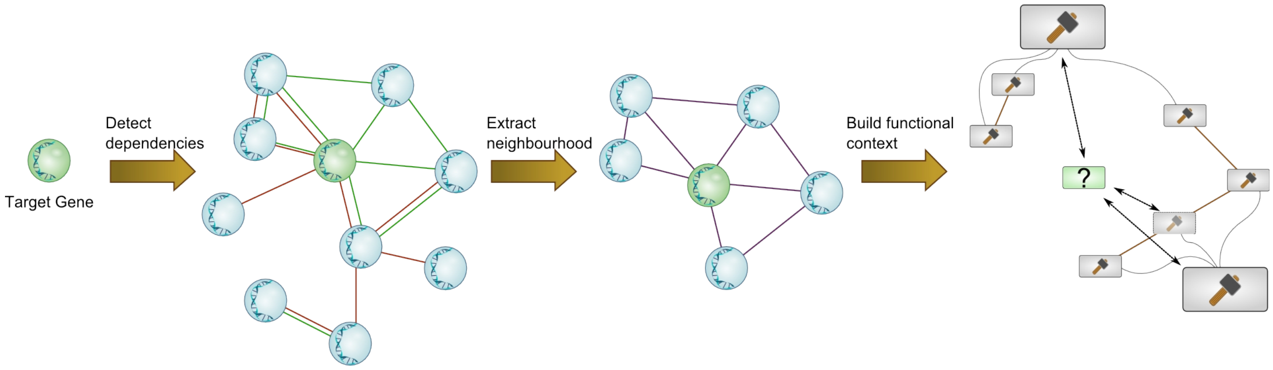


Illustration IV.5: Using functional dependence to annotate a gene

Functional inference using functional dependence follows several steps. First, functional dependency within the studied genome are established. These dependencies are then used to define a neighbourhood of the gene. The functional context formed by the functions of the other genes in this neighbourhood are then used in order to propose possible functions for the original gene.

As said, several evidence sources can attest to the functional dependence of genes or their corresponding proteins. Below, I detail those that belong to what is collectively termed "genomic context", as well as those that are "experimental data-based", along with the biological roots of these sources, and a few example methods that exploit them. I then conclude this Chapter by describing methods that call upon multiple different sources.

IV.D.2. Genomic Context-based sources

IV.D.2.a. Prokaryote genome organisation

DNA is not organised in the cell identically across all domains of life. Prokaryote genomes (*i.e.* bacterial & archaeal) are organized into cytoplasmic chromosomes and plasmids, *a contrario* to eukaryote genomes where chromosomes are secured inside a nucleus. This has led to a prokaryote-specific genome organization. Indeed, the absence of nucleus *and* endoplasmic reticulum means that transcription and translation both take place in the same cellular compartment, as is shown in Illustration IV.1: Prokaryote transcription and translation. Protein synthesis can thus be rendered more efficient by streamlining these two processes. Evolution has thus favoured prokaryote genomes on which genes coding proteins participating in same biological processes cluster together on the chromosome or on plasmids. Indeed, closely clustered genes can be transcribed and translated quickly together at the same time and place in the cell (the extreme examples being a) polycistronic mRNAs, *i.e.* mRNAs carrying several genes, and b) gene fusion, leading to one mRNA coding for two proteins). They can also share common transcription initiators and terminators, forming a structure called “operon” and thus simplifying genetic regulation of the concerned bioprocess [114]. Eukaryote genomes, with their spatially/temporally decoupled and more complex transcription and translation, rely more heavily on regulatory processes and chromosomal 3D organisation, which do not shape the organisation of genetic material in the same way. For both types of organism, “regulons” are multiple sets of genes all sharing a common regulatory mechanism that are not necessarily co-localised on the genome. Locating such structures in a studied genome gives valuable insight into its workings.

IV.D.2.b. Definition of genomic context

Many types of evidence form what is termed the “genomic context” of a gene, and are at the heart of comparative genomics [115]. The genomic context of a gene can be loosely regarded as the sum total of all data concerning the genome and other genes on this genome, that are linked mechanistically or spatially to the gene of interest. The most obvious of such links is chromosomal proximity, for which we make the underlying assumption that genes that are “close” on a genome have products that may interact in some way.

Genomic context is more easily exploitable in (though not exclusive to, as we shall see) prokaryotes, as eukaryote genomes are organized to exploit genetic regulation and genome 3D structure more than structural regulation. However, should cellular biology techniques evolve sufficiently one day to allow access to 3D representations of working eukaryote genomes, then perhaps “structural genomic context”-based methods will emerge. Progress has already been made in this direction by the characterisation of chromosomal “territories” in higher eukaryotes that appear closely linked to gene expression regulation [116].

IV.D.2.c. Gene clusters and syntenies

The foremost of the functional dependency evidences used in the study of prokaryote genomes are **gene clusters**, *i.e.* the tendency of prokaryote genes participating in a same biological process to cluster together on the chromosome, creating **operon** structures or regions bi-directionally transcribed from a central promoter. However, the definition of gene clusters in bioinformatics literature varies widely, and can be confused with that of **syntenies**, which are groups of conserved genes across multiple genomes. Here, I shall attempt to separate definitions that involve a single genome from those that involve multiple genomes.

Single-genome gene runs: The simplest form of gene neighbourhood is that of **gene runs**. These are sequences of immediately contiguous genes on a genome that respect some constraints that may vary between approaches. The most typical constraints are a) identical transcription direction (or not), and b) a maximum intergenic distance threshold. For example, the DOE IMG platform defines **chromosomal cassettes** as sequences of genes whose intergenic distance is never superior to 300 base pairs (bp), whatever the strand they are on [69], a single-genome variant of what is presented in [117,118]. The 300 base pair limit was proposed in [118]on the basis that for the set of roughly 10,000 genes from ten genomes that the authors analysed, the mean intergenic gap was 91 bp, with a standard deviation of 136 bp. As is common in statistics, a cut-off is established by taking the mean plus two standard deviations, *i.e.* 91+2*136=363, rounded down for convenience to 300 bp.

Beyond the simple notion of gene run is that of the previously-described **operon**, a structural, transcriptional and functional unit. Many methods exist that attempt to predict operons from genomic sequences, using single-genome or multiple-genome data, functional clues, and experimental evidence [119–125].

Multiple-genome syntenies: Gene clusters can be conserved (as shown by gene-gene homology relationships) across several genomes and are then called **syntenies** (*Nota bene*: the traditional meaning of the term “synteny” has, over the years, been superseded by a new meaning of the bioinformatics era, see [126], and I shall be using the latter, for want of a better term). Syntenies are even stronger evidence for functional dependence than gene runs, as high conservation indicates that the cluster structure is not a simple result of chance, but the result of evolutive pressures. Gene syntenies are generally used in prokaryote comparative genomics, though they are also studied between eukaryote genomes, where they tend to be larger supports for chromosomal rearrangements over the course of evolution (*e.g.* [127–129]). Given the drift of the term's meaning, it is not surprising that gene synteny definitions and methods of localisation are highly variable. They differ in their consideration of gene directions, of homology, in the number of genomes, in the number of allowed gene gaps... Below, I present some of the current works and their definitions that are relevant to this thesis.

Overbeek *et al.* [118,117] were the first to explicitly propose using gene clustering as an indicator of functional dependence and developed the **WIT server** to allow users to consult the calculated results of their methodology. Their gene clusters were defined as "pairs of close bidirectional best hits" between two genomes. Genes were "close" if they belonged to a same "run", defined as a sequence of genes on the chromosome along which the intergenic distance was never superior to 300 base pairs. Yanai *et al.* [130] confirmed the high-level predictive power of using such gene clusters by showing that 80% of all ortholog families whose members were in a same gene cluster in one or several genomes coded enzymes participating in a same KEGG metabolic map (see section V.A.5 for details about KEGG).

Kolesov and Frischman [131,132] propose the **SNAP** (Similarity-Neighbourhood Approach) method as a generalisation of the previous. Gene clusters are detected through the combined use of inter-genome sequence similarity and chromosomal co-localisation, basically allowing genome re-arrangements to be tracked across multiple genomes, capturing extended genomic context for a given gene. This idea was further extended in [133], presented in section VI.C.

For the IMG platform, a cassette of at least 2 genes that is conserved across at least 2 genomes (homology as defined in COG, Pfam or IMG-based families) is called a **conserved chromosomal cassette**, akin to a synteny [69]. A related definition is that of a **conserved pair of neighbouring genes** [134], which is a pair of adjacent genes whose orthologues in multiple genomes are not separated by more than a given number of intermediate genes.

A similar method is presented in [135]. Within, a suite of programs (including, but not limited to, BLAST and MCL) are used to establish gene homology. Conserved adjacent gene pairs (called positional orthologous genes or POGs for short) are then isolated, before being accreted incrementally into full-blown "synteny blocks". Results are stored in the **SynteBase** database and are made available via a java front-end called **SynteView**.

Syntenies based on the concept of gene teams over multiple genomes (groups of genes that are never separated by more than a given threshold in any of the target genomes) was used in [136] and extended for example in [137]. A domain-centred (rather than gene-centred) version of this approach is proposed in [138].

The **ADHoRe** method (Automatic Detection of Homologous Regions) [139] locates regions where the order of homologous genes is conserved between 2 genomes. The I-ADHoRe [140] compiles these 1-*versus*-1 syntenies into multi-genome syntenies.

More mathematically-precise synteny definitions exist. The **GRIMM-synteny** [141] and **Cynteny** [142] methods represent genomes as singed numbered vectors that indicate relative positions of genes and their transcription direction. The method searches for inversions, fusion/fissions and translocations in order to reconstruct whole genome rearrangements between only 2 genomes. These rearrangements can serve as the basis for determining syntenies. However, these methods are not adapted to many-to-many homology relationships in the target genomes and cannot handle insertions/deletions of genes.

Graph-based synteny definitions exist as well, wherein genes and genomes are represented as graphs, with gene neighbouring and sequence similarity-based edges linking genes. Locating syntenies becomes a question of locating structures in this graph. For example, in the Kanehisa lab (that maintains the KEGG database, see section V.A.5), "**correlated gene clusters**" are identified using an adapted version of a graph alignment algorithm [143] that they developed with another objective in mind ([144], see section VI.C.1.c for more details). The algorithm searches for groups of co-localised genes in one genome that correspond, via best hits, to groups of co-localised genes in one or several other genomes. They proceed further by clustering their correlated gene clusters into "conserved correlated gene clusters". They use all this information to construct KEGG ortholog families.

Another example is **Cyntenator** [145], which uses sequence similarity cliques to define the homology relationship (analogous to making families), and links cliques together when a significant fraction of the target genomes possess members of each clique as neighbouring genes transcribed in a same direction.

Finally, the works of [5,146] detail their own mathematical definition of a synteny (which they call a synton), along with the **CCCPart** algorithm that is described more in section VII.C.1. A mathematical generalisation of graph alignment algorithms can be found in [147]. Of these works were born the synteny-locating algorithm that is used in MicroScope, one of the main features of the platform. *In fine*, a synteny is defined in MicroScope as: a group of genes in one organism with correspondences (bidirectional best hits or at least 30% identity on 80% of the length of the shortest sequence) with a group of genes in another organism; the genes in each group many not have more than 5 intervening genes without correspondences, though intergenic distances and gene transcription frame/direction are ignored. The CCCPart algorithm also serves as a base for the CanOE strategy (see chapter VII).

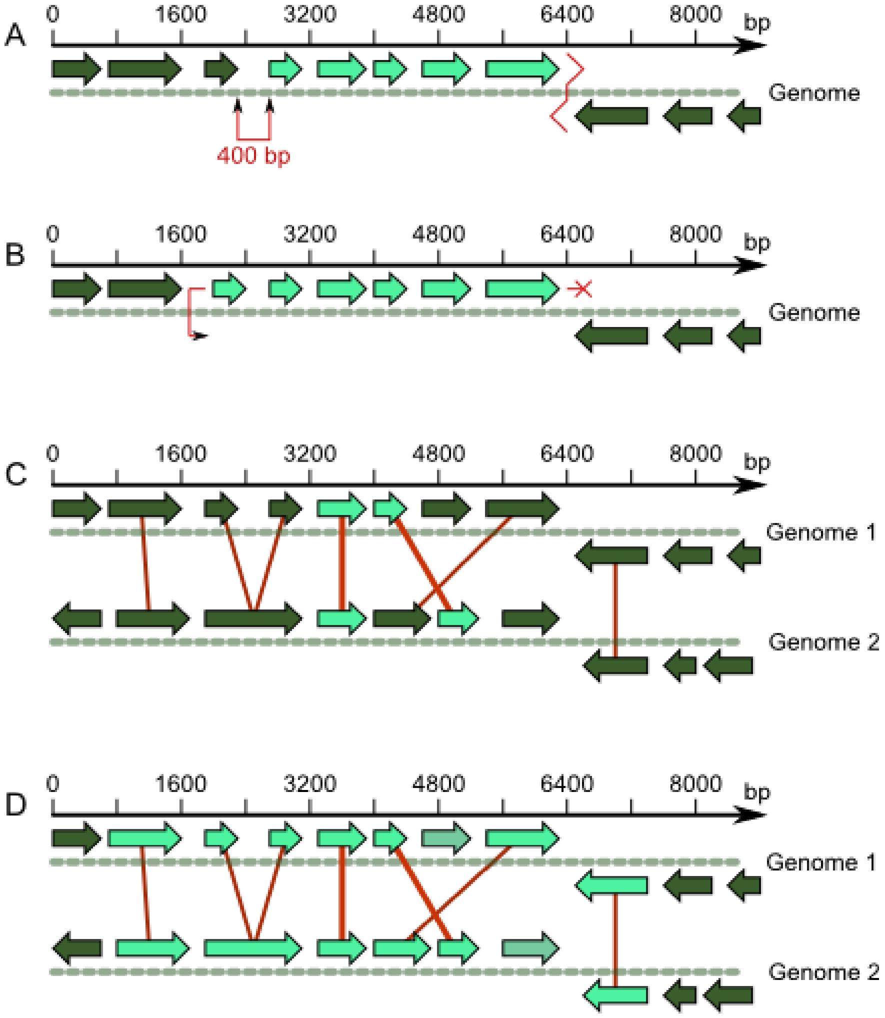


Illustration IV.6: Gene clusters and syntenies

All arrows represent genes. (A) In light green, a gene run based on a) same transcription direction (e.g. right border) and b) intergenic spacing less or equal to 300 bp (e.g. left border). (B) In light green, an operon determined by the position of a transcription initiator sequence (left) and terminator sequence (right). (C) Gene homology is represented by red edges between genomes. In light green, one gene cluster (bidirectional best hit pair) according to the works of Overbeek. Notice the requirement of homology, gene runs, the allowance for gaps, and the lack of constraint on gene order. (D) In light green, a synteny without transcription direction nor gene order constraints, allowing for a maximum of 1 consecutive gap (in gray).

All in all, many ways to determine syntenies exist, capturing various evolutive pressures, and thus not necessarily finding overlapping results. Any of them, of course, can be used as a marker of functional dependence.

IV.D.2.d. Rosetta stone (gene fusion/fission)

In some cases, chromosomal proximity is taken to an extreme when genes actually **fuse** together. A triplet of one fused gene and corresponding separate homologous genes from another genome has been called a “Rosetta stone”, as it helps decipher the interaction between gene products [112,113]. Many other works have included fusion/fission events in comparative genomics efforts [148,149] and they are now considered as a staple of genomic context. AllFuse [150] and FusionDB [151] are databases dedicated to detecting and listing fusion/fission events, with the latter extending the concept to COG families rather than just genes. Fusion/fissions are singled out in the MicroScope platform as tentative targets of interest to annotate.

IV.D.2.e. Shared regulatory sites

Genes can also be grouped together based on shared regulatory sites. These sites are usually found within non-coding DNA regions upstream of the regulated genes. Automatic regulation pattern discovery and matching software exists, such as the Regulatory Sequence Analysis Tools suite [rsat.bigre.ulb.ac.be/rsat/] developed at the BiGRe (Bioinformatique des Génomes et des Réseaux) lab [152–154]. Actually using regulatory sites in order to group genes has been carried out semi-manually in works such as [155], and gene regulation databases include RegulonDB [119,156], SwissRegulon [157] and PRODORIC [158]. Such gene groupings should be particularly close to the definition of operons or über-operons that are strong indicators of functional dependence.

IV.D.2.f. Phylogenetic profiles

Phylogenetic profiles (sometimes referred to as phylogenomic profiles) are vectors describing the presence/absence of a given gene family across many genomes. Similar phylogenetic profiles are evidence of functional dependence: two genes participating in a same biological process are more likely to be either both present, or both absent, in the same organisms, as the loss of either one would disrupt the process. [159] were the first to propose the use of gene phylogenetic profiles to measure gene dependence, and study “exact-matching-but-1” profiles to highlight a proof-of-concept case study. Many variants have since been proposed, diverging in the detection of gene orthology, the usage of binary or weighted vectors, and in profile similarity measures. Though phylogenetic profiles are still based on sequence similarity, the latter is not the source of transferred annotations, merely an integrated indicator of evolutive constraint between genes.

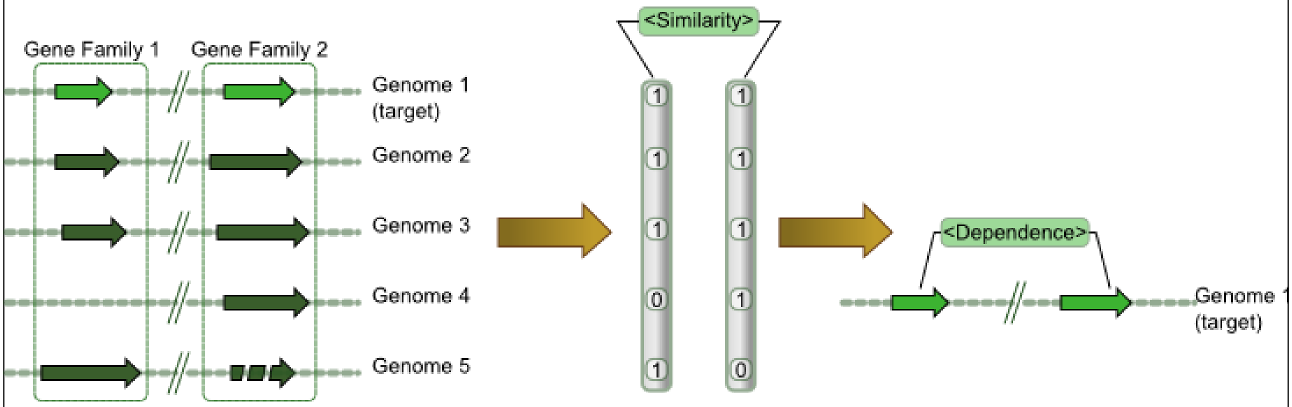


Illustration IV.7: Phylogenetic profile principle

On the left, two target genes from are target genome. Family-based homologs from other genomes are represented (one very partial match in genome 5 for gene family 1). The presence/absence across genomes is converted into binary vectors for each target gene. The two vectors are compared and a similarity metric is derived from them. This cross-genome similarity is then transferred to the target genes as a signal of functional dependence.

Phylogenetic profiles are not based on gene chromosomal proximity, and thus are genomic context indicators that can be, and have been [160] used in eukaryote genomes.

As seen, many genomic context-based functional dependency measures can be derived. Not all are equal in performances when it comes to serving as a basis for protein function prediction. The best performing of these are arguably the various versions of the gene clusters approach, as described in [161–163].

Now that we have seen gene-based genomic context, we can now proceed to functional dependency detection methods that rely on experimental data sources.

IV.D.3. Experimental data-based sources

These sources require data other than genomic data to be established. However, due to this requirement, they are not readily used across a great number of genomes, and thus did not fit in with the objectives of my thesis. I shall thus only briefly describe these sources here.

IV.D.3.a. Co-expression / co-regulation

In the same frame of mind as for shared regulatory sites, two genes presenting a similar expression profile across a set of experiments may share a common regulation mechanism, and thus be functionally related. Multiple co-regulated operons are called **regulons**.

IV.D.3.b. Co-citation

A literature-based measure of association between genes is the frequency at which both genes are cited together in the same scientific publications. Such meta-data analysis has found increasing popularity since its first uses in psychology [164,165]. Examples of using scientific literature to derive associations between genes can be found in [166,167] and in STRING (Search Tool for the Retrieval of Interacting Genes) [168].

IV.D.3.c. Protein-protein interaction

Many proteins act by interacting with other proteins, often by binding to them to form protein complexes. These physical events can be detected by wet-lab experiments such as two-hybrid techniques in yeast [169] and co-immuno-precipitation [170]. Protein-protein interaction (PPI) networks derived from such experiments have been extensively used as a different source of functional context for annotation inference [171,56]. More recently, aligning protein-protein interaction networks has been shown to improve protein orthology prediction, thus forming a better base for functional annotation transfer [172], such as with the IsoRank and IsoRank-N algorithms [173,174].

IV.D.3.d. Other data sources

Other information can be used to detect functional dependency between genes. **Growth medium-specific gene essentiality data** can give clues as to which metabolic pathways certain genes participate in [175]. **Protein 3D structure** can deliver protein-protein interaction predictions, ligand binding site predictions, or even broad fold similarities used to infer common metabolic pathways [176].

IV.D.4. Multiple-context-based annotation methods

Any of the previously cited data sources can be used individually to establish functional dependencies. However, as across the domain of bioinformatics, the trend is to **integrate** several data sources in various ways in order to improve coverage and power. The first attempts at this were **manual** and correspond to traditional bioanalysis, as in [117,177]. The first **computational** approach was in [113] where yeast gene-gene dependency edges (established from phylogenetic profiles, experimental protein-protein interaction data, sequential metabolic step catalysis, Rosetta stone fusion/fission, and correlated mRNA expression data) are basically summed together, with different confidence scores according to the different sources, in order to predict and score functional SwissProt keywords. [178] were the first to use the SVM machine learning technique to integrate gene expression data with phylogenetic profiles to predict a special functional classification in yeast. [179] used Markov Random Fields to combine physical interaction networks and functional descriptions from the Yeast Proteome Database in view of functional prediction. Joshi *et al.* [180] teach an algorithm to estimate the probability that any two genes share a similar function for several types of high-throughput data (two-hybrid screening, physical protein interactions, micro-array data, protein complexes) in Yeast. Chen *et al.* [181] train a Bayesian algorithm on yeast to transfer GO terms between proteins showing high similarities for various data sources (protein-protein interactions, protein complexes, & gene expression data). Zhu *et al.* [182] derive protein features and train a SVM algorithm that does not depend on sequence similarity to predict protein functions. Ferrer *et al.* [163] generate scores between genes reflecting the likelihood that they participate in a same biological process, on the basis of genomic context information (phylogenetic profile similarity, conserved gene neighbourhood, gene clusters, and Rosetta Stone), which they use to derive gene groups, the relevance of which is estimated by considering their conservation across multiple, selected genomes.

Over the years, many bioinformatics methods have developed by scientists of various origins (mathematicians, biologists, computer scientists, ...) in the race for the perfect protein function predictor. Presenting all of the numerous functional prediction methods that exist today is beyond the scope of this thesis. I shall thus only present a few here, and several more that can specifically answer the “orphan enzyme problem” in chapter VI. For more information on these methods, I recommend the following articles.

The review by Sharan *et al.* [56] presents several, protein-protein interaction network-centred approaches, which the authors classify into two types: those based on propagating annotations through the network, and those based on identifying clusters in the network, and propagating annotations within each cluster separately. The article by Erdin *et al.* [183] is a recent and well-written tour of protein functional prediction methods, discussing issues such as function description, and describing the main sequence-based, structure-based and network-based methodologies for functional annotation.

IV.D.4.a. Annotation platforms

All annotation platforms (such as MicroScope [4], IMG [37], the SEED [39], and ERGO [38]) formalise both sequence-based and context-based functional annotation approaches. Indeed, they offer computed results for a number of different methods concerning any target gene, such as sequence similarity, syntenies, gene clusters, phylogenetic profiles, gene/protein families, predicted domains, co-expression... The bioanalysts play the role of human integrators of this wealth of information, coercing increasingly precise annotations from the complex picture they can paint about their gene of interest.

Parts of the modus operandi of expert bioanalysts can be automated, easing their workload, focusing on interesting cases, ranking candidates. This is the true objective of all the automatic context-based methods conceived so far.

IV.D.4.b. The STRING

The **STRING** (Search Tool for the Retrieval of Interacting Genes) [184,162,149] is a database storing many different protein-protein dependency measures for proteins pairs from over 1100 organisms. Each measure is transformed into a confidence score by the following benchmarking approach:

- a true interaction is defined as a pair of proteins known to participate in a same KEGG pathway
- a positive interaction is defined as a pair of proteins with a dependence measure above a certain threshold
- establish true positive rate curve using these definitions, over all values of the given measure

- the confidence for a given measure value is the true positive rate for that value

Each confidence score can be viewed as a probability. Individual functional dependence confidence scores are combined into a single integrated confidence score by multiplying the probabilities of associations *not* predicting a functional interaction, assuming the statistical independence of each different data source.

STRING protein-protein associations can then be used manually to infer functional context and ultimately, function, for a given target gene. A relatively large-scale manual analysis of STRING results in conjunction with COG families was presented in [185]. The interface's ease of use (web query tool or access to downloadable data) and the summarising of multiple functional dependency measures into one network are a boon to bioanalysts with precise questions in mind (see also section VI.C.1.b for some relevant case studies).

Furthermore, many function prediction tools are based on STRING data, and it has become a reference source of pre-calculated functional dependence information. For example, [186] combine STRING association scores and intergenic distances to train a neural network for predicting operon structures in *Escherischia coli* & *Bacillus subtilis*. [187] develop a “network-based hierarchical Bayesian auto-probit model” for exploiting STRING association scores in order to predict GO terms for unannotated proteins.

The PINTA website [188–190] proposes a battery of prioritisation algorithms to be run on differential gene expression data populating the protein nodes of a binary STRING network in order to determine novel candidate genes, not for orphan enzymes, but rather for implication in diseases. The website is dedicated to eukaryotes (human, mouse, rat, worm and yeast). The prioritisation algorithms are Heat Kernel Ranking [189], Arnoldi Diffusion Ranking [188], Randow Graph Walk, HITS with priors [191], or k-step Markov [191].

In [192], the authors propose an algorithm to use STRING association data with yeast phenotype-associated proteins to predict phenotypes for other yeast proteins.

In [193], the authors develop a strategy of finding conserved functional modules between STRING-based binary protein-protein association networks (using NetworkBLAST [56]) to reinforce OrthoMCL-inspired GO annotation transfer between *Mycoplasma genitalium* proteins.

IV.D.4.c. Predictome

In [194], the authors describe the Predictome database, which predicts and stores functional links between genes. Links are predicted using chromosomal proximity, phylogenetic profiles, and fusion/fission. Once again, having a compiled network of dependencies relieves some of the effort that bioanalysts must put into building a representative functional context for their target gene.

IV.D.4.d. ProLinks

The ProLinks database [134] stores gene functional dependence data concerning Rosetta stone fusion/fission, phylogenetic profiles, gene clusters (groups of genes on a genome whose probability of being in a same operon is a function of intergenic spacing), gene neighbours (gene pairs conserved across genomes without being separated by more than a certain number of genes), and literature co-citation. Each of these methods generates a gene interaction p-value. In order to be comparable, these p-values are scaled into probabilities by a COG-based benchmarking approach.

Prolinks dependencies are exploited by the “Pathway Hole Filler - Genomic Context” algorithm that is detailed in section VI.C.1.e.

IV.D.4.e. GeneMANIA

GeneMANIA [195,196] is a web server that dynamically interrogates a database comprising many gene-gene/protein-protein interaction networks of various sources (co-expression, physical interaction, genetic interaction, co-localisation, shared pathway, shared domains, etc.) with a query set of seed genes. Local functional association networks from each data source are combined using dynamically-calculated weights, and gene function labels are propagated across the integrated network and can be used as predictions for the seed genes. GeneMANIA currently supports 6 eukaryote organisms (*Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophilia melanogaster*, *Mus musculus*, *Homo sapiens*, and *Saccharomyces cerevisae*).

These annotation tools allow the study of the functions of prokaryote (and for some, eukaryote) genes. One category of functions is of particular interest to biologists, biochemists and biotechnologists: metabolic enzymatic activities. As the goal of this thesis is the development of methods for annotating genes with metabolic reactions, I shall now briefly present the basics of prokaryote metabolism.

V. Metabolism

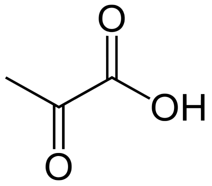
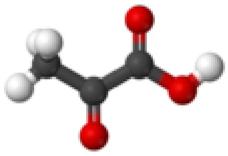
V.A. Key Actors

V.A.1. Compounds

Chemical compounds are made up by several atoms of one or more chemical elements (e.g. carbon, oxygen, hydrogen...) that are maintained in a more or less rigid three-dimensional structure by chemical bonds. Some compounds can be small, others very large with many atoms (e.g. DNA, or crystalline structures). Compounds are called **metabolites** when they correspond to small molecules or ions that are processed in biochemical (metabolic) reactions (see section V.A.2).

Compounds are identified primarily by chemical formulae that describe the number of atoms of each element that compose it, with various systems for representing various levels of substructure detail; the simplest of these is the **general chemical formula**. The three dimensional structure itself can be described by a **structural formula**, using a variety of representation systems, the most common in biology being the skeletal formulae which supposes carbon-based molecular backbones with implicit hydrogenation. Attempts have been made in order to make chemical structure description easily accessible to humans and computers. For example, the SMILES (simplified molecular input line entry specification) can describe any molecule as a linear series of characters that represent successive atoms (less hydrogen) along the molecules' backbone, all cycles being “broken”. The InChi scheme (IUPAC International Chemical Identifier) also proposes linear text descriptions of molecules that deal more directly with the chemical structure. Additionally, InChi defines a standard for hashing (a computational approach to compressing data) in order to shorten the representation.

As an example, several different representations of a compound central to metabolism across all domains of life are illustrated in the following table.

Common name	Pyruvic acid
General formula	$C_3H_4O_3$
Structural formula (hydrogen atoms hidden, carbon atoms implicit)	
Balls-and-sticks 3D representation	
SMILES formula	O=C(C(=O)O)C
InChi formula (non-hashed)	1S/C3H4O3/c1-2(4)3(5)6/h1H3,(H,5,6)

V.A.2. Reactions

Chemical reactions involve the transformation of one or more substrate compounds into one or more product compounds (collectively called reactants). This transformation can be described at its simplest by a stoichiometric formula which compiles the number of molecules of each involved compound and the reaction direction. More detailed descriptions exist that detail the chemical structures of the reactants and the mechanism of the transformation. Chemical reactions are generally spontaneous and reversible (*i.e.* can take place in both directions, from substrates to products or from products to substrates) but not all reactions nor directions are equal from an energetic point of view, and most are extremely unlikely to happen spontaneously. An example reaction is given in the following figure:

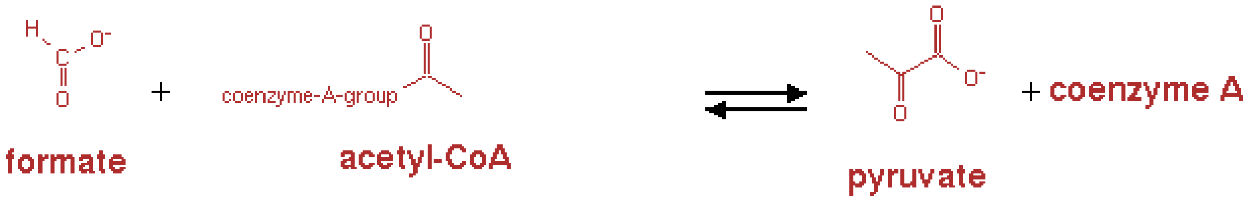


Illustration V.1: Example reaction for pyruvate

A metabolic reaction taken from MetaCyc [www.metacyc.org], illustrating the reversible transformation of pyruvate (the base ion for pyruvic acid) into formate.

The speed *v*+ at which a chemical reaction transforms substrates into products (or *v*- for products into substrates) is dependant on a number of factors:

- **reactant activities**, which are closely linked to reactant concentrations for entities en aqueous solution
- **reaction surface area**, or the sum total of all reactant-reactant interface surfaces, of particular importance when reactants do not belong to a same phase (*i.e.* solid, aqueous, alcohol, gaseous...)
- **pressure** and **temperature**, which influence how often reactants can approach each other sufficiently for the transformation to take place
- presence of a **catalyst**
- **activation energy** E_A, which is the minimum amount of energy that must be provided to reactants for a reaction to initiate.

Given initial concentrations of reactants in aqueous solution, a chemical reaction will modify these concentrations over time until they reach asymptotic limits that correspond to chemical equilibrium, a sign that *v*+ and *v*- are equal. The **equilibrium constant** K_{eq} is the ratio of the reaction speeds.

V.A.3. Enzymes

Many reactions are considered irreversible in typical laboratory operating conditions (25°C, 1 atmosphere pressure) and many require specific conditions (*e.g.* solution acidity) or an energy source (*e.g.* heat) to take place. The use of **catalysts**, *i.e.* substances that can participate in the reaction but that are not modified durably by it, can greatly increase the efficiency of a given reaction. Catalysts reduce the activation energy a reaction requires to initiate, effectively increasing the reaction speed and/or reducing exterior energy source requirements, without affecting the equilibrium constant. In inorganic chemistry for example, the use of platinum as a catalyst greatly reduces the heat and pressure conditions required for organic molecule hydrogenation to take place.

Metabolic reactions are chemical reactions that take place in a living cell. High levels of heat and pressure are obviously incompatible with Life (even thermophile bacteria, that can withstand temperatures close to the boiling point of water, could not withstand the high temperatures required to hydrogenate unsaturated fatty acids for example). Hence, biologically-available catalysts are of prime importance in the biochemistry of life.

The great majority of biocatalysts are **enzymes**, which are a particular type of protein. Some other biological molecules have catalytic properties, such as ribosomes, ribozymes, and other types of active RNAs. Unlike usual catalysts, however, enzymes may have particular behaviours in respect to the equilibrium constant. Indeed, if a given constant heavily favours one reaction direction, then conditions (such as sequential enzymatic reactions) may ensure that the enzyme *effectively* catalyses only this direction. Furthermore, enzymes often couple secondary reactions to their primary ones. **Enzymatic coupling** allows the transfer of chemical energy between reactions, *e.g.* using ATP hydrolysis to initiate liposynthesis reaction, or using the concentration-gradient driven passage of protons through a membrane ATP synthetase in order to regenerate ATP.

Enzymology, or the study of enzymes, generally assumes that reactants and the studied enzyme are in an aqueous solution and at conventional laboratory conditions compatible with Life. However, studying enzymatic activities is more complex than for simpler chemical reactions, as the catalyst intervenes actively in the reaction mechanism and often couples one reaction to another, rendering the depiction of a molecular model difficult. The simplest of such models is the **Michaelis-Menten model**, that considers that the enzyme and its substrates must bind in an easily reversible step into an enzyme-substrate complex, that is then transformed into a product by a more favourable second step. Closed-form mathematical descriptions of reactant and free-enzyme concentrations, as well as reaction speeds, over time are derivable with this model, and several quantities of interest are used to characterise enzymatic reactions:

- Unlike inorganic catalysts, enzymes have saturation dynamics; indeed, each enzyme has only a limited number of active sites (usually one per reaction type except for multimeric enzyme complexes). This means that the reaction rate approaches an asymptotic maximum, noted **V_{max}**, as the concentration of substrate increases.
- The Michaelis constant **K_M** is the substrate concentration at which the global reaction rate is at half its maximum V_{max}, and is a measure of the inverse of the affinity of the enzyme for the substrate. The smaller the K_M, the higher the affinity, and the faster V_{max} will be reached.
- V_{max} can be expressed as a function of the initial quantity of free enzyme times a constant. This constant is the turnover number, **k_{cat}**, which is the maximum number of substrate molecules that one enzyme molecule can convert per second. The enzymatic efficiency, K_{cat} / K_M, is useful in comparing enzymes.

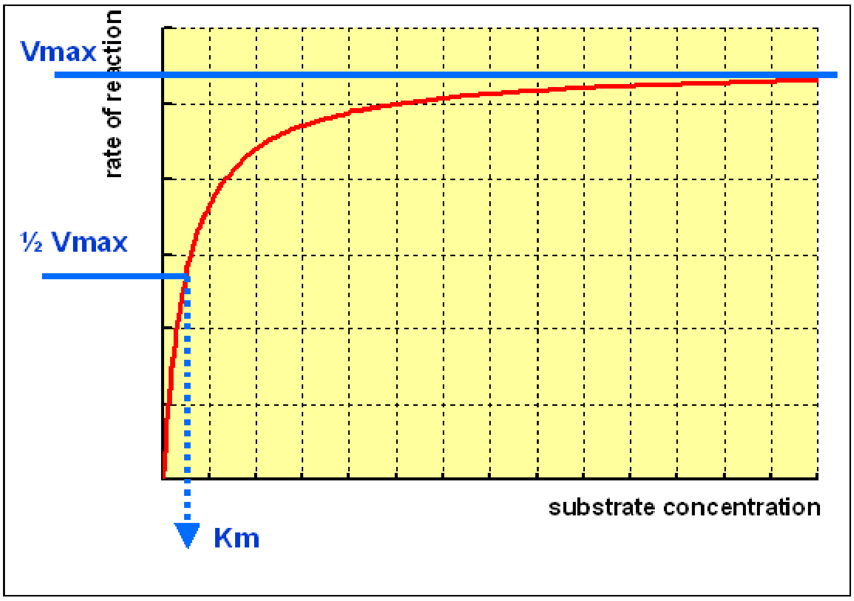


Illustration V.2: Experimental determination of Vmax

The maximum reaction rate of an enzyme that follows Michaelis-Menten reaction kinetics can be determined by plotting the reaction rate (*i.e.* the speed at which substrate is consumed or products produced) of the enzyme against different concentrations of substrate. The reaction rate approaches an asymptotic maximum for increasing substrate concentrations. The Michaelis constant K_M can also be read from the same plot, as shown here.

Protein activity is heavily dependant on its 3D structure and amino acid sequence. Indeed, an enzyme intervenes in the chemical reaction thanks to several mechanisms such as correctly positioning the substrates, distorting electron fields, forming temporary chemical bonds, readily providing labile protons, etc., thanks to the disposition of its amino acid side chains (“residues”). Structural proteins stick together thanks to specific binding sites. Regulatory proteins interact with other proteins or molecules thanks to 3D domains. Actually, the entities coexisting within the bubbling cauldron of a cell are so numerous and varied, that many different types of interactions can take place between them, leading to a phenomenon known as **functional promiscuity** (see section V.C.1). Thus, the relationship between coding gene and activity can be a many-to-many one. For example, some enzymes are made up of one or more identical (different) proteins, and are thus homopolymeric (heteropolymeric, respectively). Inversely, one activity may be catalysed by different proteins without evolutive relationships, which are then called **isozymes** or isoenzymes.

All in all, this serves to show that *any* functional annotation can always be considered with doubt, as it can not only be possibly wrong or biologically irrelevant, but it can be incomplete in many unexpected ways, especially since experimentally testing every different protein for every function imaginable in each different type of cell is simply not feasible. This should always be kept in mind when creating functional annotations or when studying the activity of a given enzyme, even when its coding gene's context points rather clearly to a specific activity, such as a step out of a metabolic pathway.

Enzymes catalyse biochemical reactions, transforming environment substrates into molecules better adapted to the cell's requirements. A single substrate can be submitted to sequential transformations, gradually modifying or incorporating it. Such chains of transformations are known as metabolic pathways and are presented below.

V.A.4. Metabolic pathways

Metabolic reactions do not operate alone within the cell; indeed, often the product of one reaction is a substrate of another, allowing a cell to transform one metabolite into another in a series of simple steps. A **metabolic pathway** can be broadly defined as a set of linked metabolic reactions that participate in a same higher-level metabolic process, along with contextual and experimental data. However, the exact definition for what a metabolic pathway actually is remains in debate and some of the important points are discussed below.

Historically, metabolic pathways were described experimentally by biochemists in a given organism. Metabolic pathway definition was thus organism-specific and its reactions respected stoichiometry balance. Phenotypic data, gene regulation data, gene essentiality data and genomic data could also be used to describe environment-specific pathways, their regulation and genomic localisation, such as with the lactose operon [197]. Most importantly, certain compounds from the pathway were considered as more important than others: "**main compounds**" captured the most important path through the metabolic reactions, not linking reactions via secondary compounds (often that could be considered to come from an available cellular "pool", such as with energetic substrates). These paths were considered biologically relevant as they followed the flow of biochemically relevant atoms (such as carbon and nitrogen) and could thus be experimentally validated by atom-tracing experiments (*e.g.* [198]). **Anabolic** pathways concerned the synthesis of compounds from others, usually at the expense of energetic substrates (*e.g.* DNA replication); **catabolic** pathways degrade compounds into simpler products and energetic substrates (*e.g.* glycolysis).

V.A.5. Metabolic Resources

V.A.5.a. Compounds

The Chemical Entities of Biological Interest (ChEBI) is a publicly-accessible database dedicated to describing chemical atoms or compounds that are known to participate in the metabolism of living organisms. It is maintained by the European Bioinformatics Institute (EBI) (located in the UK). Large molecules (such as biopolymers) are not included in ChEBI. The ontology used in ChEBI conforms to directives from both the International Union of Pure and Applied Chemistry (IUPAC) and the International Union of Biochemistry and Molecular Biology (IUBMB). PubChem is another public database maintained at the National Center for Biotechnology Information (NCBI) (located in the USA) that describes molecules, complexes and mixtures. Both of these resources allow retrieval of various chemical data, as well as keyword or chemical structure searches.

V.A.5.b. Reactions and Enzymes

The International Union of Biochemistry and Molecular Biology (IUBMB [www.iubmb.org]) provides community interactions and statutes on values, ethics and standards in biochemistry-realted scientific domains. Amongst other things, they propose the **EC (Enzyme Commission) classification** scheme [www.chem.qmul.ac.uk/iubmb/enzyme], which is a vocabulary specific to metabolic reactions that are catalysed by enzymes (*i.e.* enzymatic activities). Distinct activities are identified by a 4-digit number. Each number refers to a increasingly precise biochemical description. The first number, for example, describes a “reaction group type” and can be any number from 1 to 6 (1: oxidoreductases, 2: transferases, 3: hydrolases, 4: lyases, 5: isomerases, and 6: ligases). The 4th digit generally refers to substrate specificity. EC numbers present several inconveniences, of which: sub-class digits do not correspond between classes; the sequence-structure-EC number relationship is not straightforward; and EC numbers are not designed to take into account multi-functional enzymes. All of these are well-known problems for computational biology [100,199,200,183].

Quite obviously, EC numbers cannot be used for describing non-metabolic functions.

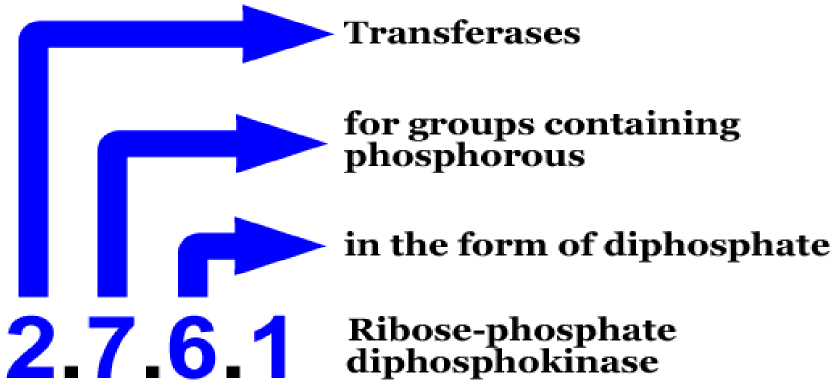

Transferases
for groups containing phosphorous
in the form of diphosphate
2.7.6.1
Ribose-phosphate diphosphokinase

Illustration V.3: The Enzyme Commission number principle

Distinct activities acknowledged by the IUBMB are identified by a 4-digit number. Each number refers to a increasingly precise biochemical description. The first number describes a “reaction class”. The 2nd describes a “reaction sub-class”. The third gives an indication on the involved compounds. The 4th digit is arbitrary, and can be considered to refer to substrate specificity.

Dedicated databases exist that census enzymes and associated information. Some sequence databanks generically track which genes/proteins are associated with which enzymatic activities, *e.g.* UniProt [201]. The ENZYME database catalogues EC numbers that are associated to proteins in UniProtKB/Swiss-Prot, guaranteeing high-quality annotations [202]. The BRENDA database [203,204] is dedicated to compiling experimental biochemical information for each EC number in each organism, such as the enzyme reaction constants presented previously in this section, as well as bibliographic references that back them. Rhea [http://www.ebi.ac.uk/rhea/] is the EBI's academic metabolic reaction database. However, many databases describing metabolic reactions also put them into biochemical contexts: metabolic pathways.

V.A.5.c. Metabolic pathways

Several bioinformatics resources exist that formalise metabolites, metabolic reactions and metabolic pathways in their own way.

KEGG: The Kyoto Encyclopedia of Genes and Genomes (KEGG) is, despite its name, first and foremost an academic database of metabolic reactions and pathways, with large pan-organism metabolic maps with organism-specific projections [205]. A KEGG map is made up of KEGG Reactions and KEGG Compounds. KEGG Reactions generally correspond to EC numbers, though many KEGG reactions are specific instances of generic EC numbers. KEGG supports the KEGG RPAIR database, which contains ReactionPairs. These are -in essence- pairs of compounds involved in a given reaction, with focus upon chemical groups that are modified or transferred during the reaction between molecules. The KEGG LIGAND database is an integrated database containing (amongst other things) KEGG Reactions, Compounds, and RPAIRs. Other more specific resources are available, such as KEGG Glycan and KEGG DRUGS. To complement the large KEGG maps, a smaller-scale definition of metabolic pathways is implemented in the KEGG modules resource.

EcoCyc, MetaCyc and BioCyc: Peter D. Karp et al. developed the EcoCyc database in 1997 [206,207]. EcoCyc comprised an original database system storing experimentally-demonstrated information on the genes, enzymes, and associated metabolic reactions and pathways found within the *Esherichia coli* K-12 genome, as well as a graphical interface for accessing the data. Since, the same procedure has been applied to the genomes of many well-studied prokaryote and eukaryote organisms, creating what have become known as organism-specific pathway/genome databases (PGDBs) for each. The collection of databases with experimental results is called MetaCyc, and together with EcoCyc they form the “tier 1” PGDBs. Building on the success of EcoCyc and MetaCyc, other *Cyc databases were created, using bioinformatics predictions benefiting from the previous experimental demonstrations in other organisms. These predictions were established by the Pathway Tools software suite [208,209], and the generated PGDBs are “tier 2” and “tier 3” according to the level of manual verification and curation they have undergone since. The MicroCyc component of the MicroScope platform is an extended pathway genome resource stemming from a “local” installation of the MetaCyc database and software. It contains the PGDBs of all the MicroScope organisms.

Reactome: Reactome [210–212] is an open-source expert-curated database of human reactions and pathways with multiple inter-database cross links, with a focus on inferring orthology events in other higher eukaryote species.

Unipathway: Unipathway [http://www.grenoble.prabi.fr/obiwarehouse/unipathway] is an expert-curated database with a novel reaction-chain model. It also stores metadata about its hierarchically-organised pathway definitions. It is integrated with the UniProtKB resource [213].

ExPASy: ExPASy provides a web access to the Roche Applied Science "Biochemical Pathways" wall charts [http://web.expasy.org/pathways/]. The portal allows searching by keywords or EC numbers, and returns clickable pathway images with links to the ENZYME database. There is, however, no underlying data model allowing the resource to be queried in respect to the pathways.

A survey of the metabolic resources, classified into “resource families” such as MetaCyc, KEGG, Reactome and BiGG, can be found in [214], where the authors compare present data types, metabolism coverage, and analysis tools. During this thesis, I worked almost exclusively with KEGG and MetaCyc, due to their availability in the MicroScope platform. See [Illustration V.4: The allantoin degradation pathway in KEGG and MetaCyc] in the following section for a graphical representation of a pathway from each.

V.B. Dealing with Metabolic Pathways

V.B.1. Metabolic pathway representations

V.B.1.a. Representing pathways as networks

With the advent of computational biology arose the need to represent metabolic pathways using abstract, mathematical models that computers could process. A common and humanly-readable way of representing one or several metabolic pathways is as **networks** (also known as **graphs**). Three types of metabolic pathway network representation exist:

Compound-centred network: network nodes are chemical compounds, and directed network (hyper-)edges are the reactions that can transform the compounds. The use of hyper edges (*i.e.* edges with multiple start and end vertices) for reactions modifying several metabolites at once is useful for capturing reaction reversibility, though it implies a rather daunting network complexity. Hyperedges can be broken down into “normal” edges, though in this case a single reaction would correspond to multiple edges (and a same edge could correspond to multiple reactions).

Reaction-centred network: network nodes are the reactions, with network edges added between two reactions when they share a common compound. This representation is the least humanly readable, as the conceptual link from a compound of interest to the clique (see graph theory annex) of reactions it links in this sort of graph is not easy nor readily visible.

Bipartite network: both reactions and compounds are represented by different types of nodes, and edges indicate compound participation in linked reactions. This representation is the most understandable, but the existence of two types of nodes increases algorithmic & computational complexity, as well as posing reversibility tracking problems. KEGG and MetaCyc use this kind of representation, even if the underlying data models are more complex (see [Illustration V.4: The allantoin degradation pathway in KEGG and MetaCyc] on the next page for a graphical representation of each).

Petri Network: a Petri network is a special kind of bipartite graph composed of two types of nodes called *places* and *transitions*, linked together by directed *arcs*. *Tokens* can occupy the places, and can be moved about via the arcs when transitions *fire*. They were initially designed to describe chemical reactions, but are also suited to metabolic networks. In the latter context, places represent types of molecules, and tokens are instances of these molecules, that are transformed by metabolic reaction transitions. A seminal paper that presents the methodology is [215].

Other ways of modelling metabolic networks exist, though it is not necessary in this thesis to develop them all. Do note that it is possible to pass from one network representation to another, if the necessary information is available, which is particularly useful for producing human-readable results. In the rest of this manuscript, we will use compound, reaction or bipartite compound/reaction graph representations. For each of these, several choices must be made on how to process metabolic information.

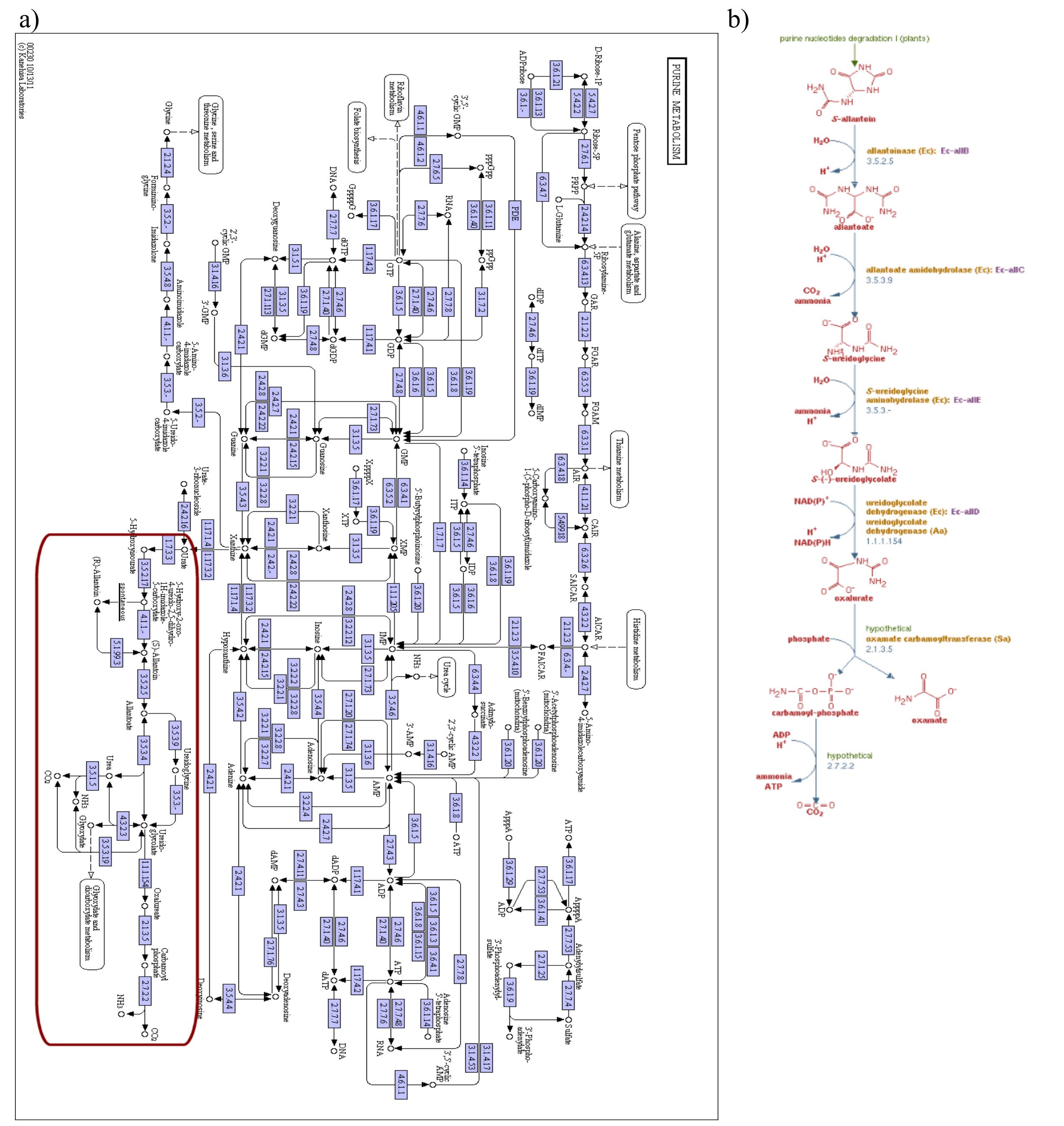

a)
PURINE METABOLISM
00230 10/13/11
(c) Kanehisa Laboratories
b)
purine nucleotides degradation I (plants)
S-allantoin
H2O
H+
allantoinase (Ec): Ec-allB
3.5.2.5
allantoate
H2O
H+
CO2
ammonia
allantoate amidohydrolase (Ec): Ec-allC
3.5.3.9
S-ureidoglycine
H2O
ammonia
H+
S-ureidoglycine aminohydrolase (Ec): Ec-allE
3.5.3.-
S-(-)-ureidoglycolate
NAD(P)+
H+
NAD(P)H
ureidoglycolate dehydrogenase (Ec): Ec-allD
ureidoglycolate dehydrogenase (Aa)
1.1.1.154
oxalurate
phosphate
hypothetical
oxamate carbamoyltransferase (Sa)
2.1.3.5
carbamoyl-phosphate
oxamate
ADP
H+
ammonia
ATP
hypothetical
2.7.2.2
CO2

Illustration V.4: The allantoin degradation pathway in KEGG and MetaCyc

a) What biochemists consider as the anaerobic degradation of allantoin pathway is, in the KEGG database, part of a broader metabolic map called “Purine Metabolism” (ec00230). The allantoin degradation part is surrounded by a red box. b) A dedicated pathway object exists in the MetaCyc database to describe this metabolic process (PWY0-41).

V.B.1.b. Metabolic network scope

First of all is the **scope**: a) does the network represent metabolic data for one, or for all organisms ? and b) does the network represent a single metabolic pathway, or the sum of all metabolic pathways ? One network of note is the **Global Metabolic Network**, representing all metabolic reaction connectivity across all known organisms.

V.B.1.c. Metabolic network edge building

The second important choice is that of which compounds are used to link reactions together. As previously mentioned, historical metabolic pathways define main compounds and secondary compounds. Linking reactions by ubiquitous compounds such as water (a common product) or Adenosine Tri-Phosphate (ATP, a common chemical energetic substrate) results in highly connected networks which are not necessarily relevant, either biologically, computationally or graphically. For example, when representing the glycolysis pathway, which describes the general degradation of glucose and the use of the liberated energy to produce ATP, the pathway should represent the fate of the atoms of most important pathway-relevant compounds, namely glucose itself. It is natural to link the successive reactions that transform glucose; it would not be natural to link together all the different reactions from the pathway that actually produce (or consume) ATP: as an energetic substrate, this would be biologically irrelevant, even if *in situ*, ATP produced by one enzyme could very well be consumed by another in the pathway. Conversely, ATP can be a main compound in certain metabolic pathways, such as in purine metabolism, where it is a substrate for RNA and DNA synthesising processes.

Different computational biology protocols have been developed to address the problem of identifying main compounds in metabolic networks:

- connect reactions by all shared compounds
- use main compound definitions where they are available
- attempt to infer main/secondary compound assignments

The first case is trivial, leading to heavily connected networks, and generally requiring some sort of filtering of results found with them. The second case relies on the definition of main compounds that can be extracted from experimental metabolic pathway definitions. The last case relies on predictive methods such as:

Compound removal: a heavily-used (ubiquitous) compound should be either removed or at least less favoured as it is less likely to be pathway-relevant.

Computational atom tracing: attempt to match computationally the atoms from the structures of a reaction's substrates and products; use this knowledge over several successive reactions to decide which compounds are “main”.

Compound removal on the basis of degree is commonly used, *e.g.* in [216]. Using weighted networks that impose a heavier cost of passing through ubiquitous compounds has since been shown to outperform previous protocols and has been used for example in [217]. Compound matching corresponds much more to the biological point of view on metabolic networks (*i.e.* focuses on the fate of key atoms or molecular substructures) and has been extensively explored over the past decade [218–224].

V.B.2. Metabolic pathway reconstructions

In order to build a comprehensive global view of an organism's metabolism, it has become commonplace to carry out a “genome-scale metabolic reconstruction”. This is a computational network model of the chemical transformations that can take place within a cell, extracted and extrapolated from the sum total of metabolic knowledge with which its genome is annotated. Metabolic reconstructions can then be used to complete current metabolic knowledge, to align observed phenotypes with current metabolic knowledge, to compare the metabolisms of different organisms [225], to make metabolic predictions, to help with metabolic engineering, to help understand the evolution of metabolism, *etc* [226].

V.B.2.a. Protocols

A metabolic reconstruction starts with the extraction of all known metabolic reactions assigned to the genes of the target genome. After this, two strategies are possible.

The first strategy is the *ab initio* reconstruction of the metabolic pathways from known activities. The latter serve as anchor points, between which additional reactions can be added so that all reactions form a connected component. The choice of additional reactions can be based on connected reactions from a global metabolic map of all known metabolic activities, or can be based on inferring required chemical transformations. For example, Boyer *et al.* [221] reconstruct pathways on the basis of atom transfers. Faust *et al.* [227] walk a global metabolic network with nodes weighted in order to favour non-ubiquitous compounds so as to connect multiple known reactions. A review of such methods can be found in [227,226].

The other strategy is reconstruction by homology, where pathway existence is extrapolated from present reactions and known pathways in other, phylogenetically close organisms. This is typically what the Pathologic software of the Pathway Logic Software Suite does [209]. Given an initial genome annotation file (such as a GenBank file), PathoLogic first establishes the list of all known metabolic reactions in the genome using annotated EC numbers and/or product descriptions, before assessing whether there is sufficient evidence for the existence of each MetaCyc pathway within the studied organism[8]. KEGG, however, does not benefit from such top-down reconstruction; their organism-specific pathway maps are merely projections of current knowledge onto global maps.

In both cases, inconsistencies (mis-annotations, dead-end metabolites, etc) can be checked for and manually curated, iteratively improving the network [228].

V.B.2.b. Reaction gaps

With a metabolic reconstruction, dead-end metabolites (*i.e.*, a metabolite in a metabolic network that is either only consumed or only produced) and incomplete metabolic pathways become apparent. Reactions that are missing but are required to complete the map are known as **reaction gaps** and are often considered as potential metabolic functions that require uncovering in the genome of an organism. In methods such as [216], reaction gaps derived from metabolic reconstructions are specifically targeted for the creation of novel functional annotations (see section VI.C.1.d for more detail).

V.C. Biocatalysis applications

V.C.1. Enzyme promiscuity

Previously-discussed protein promiscuity also applies to enzymes. Enzyme-specific promiscuity has generally been classified into two types:

- **Substrate promiscuity** describes the fact that some enzymes may catalyse the same chemical reaction but on multiple similar substrates; the opposite of this is high substrate specificity. Both of these cases have been observed by biochemists.
- **Catalytic promiscuity** refers to the ability of an enzyme to catalyse different chemical

[8] Arbitrarily, predicted pathways are declared “possibly” present when their completion (*i.e.*, the ratio of the number of present reactions from the pathway over the total number of reactions in the pathway) is in]O%; 50] ; they are “probably” present in]50%; 100[.

reactions. This is often observed for multi-functional enzymes or protein complexes that catalyse successive steps from a given metabolic pathway.

Comparing relative enzyme efficiencies K_{cat} can indicate which substrates/reactions are the most important for the cell (as it is expected to be submitted to more stringent selection which should render the enzyme more efficient for it), though *in vitro* and *in vivo* conditions might be different enough to scramble the signal, or substrate preference can be context-dependant [229]. It is even thought that entire metabolic pathways could actually present yet-unknown alternative activities that would be relevant only under specific conditions; this is called "underground metabolism" [230] and is thought to be behind much of the "dark matter" of modern metabolomics [229].

V.C.2. Industrial applications

Using the chemical transformation properties of living organisms has been an age-old practice in human culture. Be it for fermenting hops into beer, grapes into wine, starch into raised bread, or fixation of leather by saliva enzymes, humans have exploited organic substances throughout history, without even knowing exactly what they were doing until the 19th century. As biological sciences progressed, the origin of biologically catalysed transformations became apparent, and enzymes started to be isolated from parent organisms in view of scaled-up applications. However, the lack of biological knowledge, know-how and the inherent complexity of life constrained biochemical efforts. Since the 1970s, these hurdles have been jumped, thanks to genetic engineering and molecular biology, opening new possibilities for industrial developments [231].

The advantages and drawbacks of biochemical over chemical transformations are numerous and cannot all be listed here. Of particular note to the non-initiate are (amongst others):

Advantages: enzymes often function at conditions not harmful to life, and are thus much easier to manipulate; they are very efficient catalysts and can be used in small quantities; they are entirely biodegradable, with no impact on the environment, which is particularly interesting in this day and age of green technologies

Drawbacks: enzymes can be fragile; enzymes can be hard to isolate in large quantities; enzymes can depend on co-factors that require regenerating (such as nicotinamide adenine dinucleotide, NAD^+, a common enzyme cofactor involved in many redox metabolic reactions); enzymes are often kinetically challenged by substrate/product concentrations; and the full spectrum of possible enzymatic transformations is not yet known.

It is in particular this last point that the methods explored in this thesis might help to alleviate.

Our knowledge of prokaryote (or even eukaryote, for that matter) metabolism continues to be plagued by many gaping holes. The works in this thesis hope to address at least part of this problem by helping scientists to comprehend the inner workings of prokaryote organisms, which as pointed out might give rise to interesting industrial applications. In the following Chapter, I shall focus on one specific issue of our knowledge of metabolism: the orphan enzyme. I will present the extent of this problem, as well as several existing methods that strive to address it.

VI. Orphan enzymes and how to adopt them

VI.A. Definitions

Diverse biochemical experiments can reveal metabolic enzymatic activities in cultivated organisms. However, actually associating gene/protein sequences with these activities is a tedious and costly process (see chapter IV). As a consequence, such associations lag behind the discovery of new activities, leading to an accumulation of “metabolic knowledge holes” in our representations of metabolism.

Two types of “holes” are formalised in [232], depending on the state of current metabolic knowledge. On the one hand, when either the presence of an incomplete metabolic pathway, or the existence of dead-ends in a metabolic network reconstruction, suggest that a given organism should be able to catalyse a metabolic reaction (for the pathway to be complete), but the gene for the corresponding enzyme is unknown, then this organism is said to have a “**reaction gap**”. On the other hand, if an organism is known (thanks to experimental evidence) to produce an enzyme catalysing a given reaction, but the gene for it is unknown, then this metabolic activity is called an “**orphan reaction**”[9]. Due to the correspondence between gene and reaction transiting via an enzyme, these orphan reactions are sometimes abusively nicknamed “orphan enzymes” for short. So as to be as precise as possible, I will employ the term “orphan reaction” to designate a “**local sequence-orphan enzymatic activity**”, where “local” indicates that the activity is a sequence-orphan in the studied organism. The reasons for this distinction will become clear in the paragraph after next.

Establishing the list of organisms able or unable to catalyse a given reaction is not a trivial endeavour. Indeed, resources such as BRENDA [203] catalogue known activities for each organism, but not their absence. Correctly establishing per-organism orphan enzymes would require both types of knowledge. For this reason, a common work hypothesis is that all organisms can potentially code all known activities, until the perfect annotation of their genome proves otherwise. Hence, orphan enzymes can be supposed to exist even in organisms with no experimental evidence that they can catalyse them (or not).

Building upon this hypothesis, it becomes possible to define phylogenetic **scopes** for orphan enzymes. An enzymatic activity that is orphan in a single organism is a **local orphan reaction**; if it

[9] Not to be confused with “ORFans”, *i.e.* Open Reading Frames with no detected homology to any already-annotated ORFs.

is orphan across a given clade (*i.e.* across a single branch of the Tree of Life), it is a **clade-specific orphan reaction**; finally, if it is orphan across all sequenced genomes, it is a **global orphan reaction**. An example of clade-specific orphan enzymatic activity is EC 1.1.1.10 (L-xylulose reductase), which has known coding genes in several Eukaryotes (including various fungi, *homo sapiens*, *mus musculus*...), and although it is known to be catalysed in *Erwinia* sp., no prokaryote genes encoding it are known: it is thus a prokaryote orphan reaction.

In the rest of this work, I will work with the hypothesis that all organisms could catalyse all reactions, and will restrict myself to prokaryote orphans unless otherwise specified.

The main objective of methods aiming to solve the orphan enzyme problem is to find **candidate genes** for the target activity. Candidates should preferably be ranked according to some measure of plausibility.

Another issue in bioinformatics uses the term "candidate genes". In genetic studies, scientists attempt to discover which genes participate in a given phenotype or disease on the basis of experimental evidence. As for the orphan enzyme problem, the objective is to identify these genes, evaluate and rank the likelihood of their involvement in a higher-order process.

VI.B. Adoption status

The high number of global sequence-orphan enzymatic activities, and their impact on the performances of automatic gene annotation methods, has motivated the call of several scientists for a "global enzymatic genomics initiative" [233–236]. Several surveys of enzyme orphans have been undertaken, trying to establish exactly how many orphan enzymes there are, and whether they are averred orphans, or merely artefactual due to lack of proper efforts of scientists submitting results to journals or to biological databases [2,237]. In [2], Karp shows that almost 18% of enzymatic activities are probably artefactual (fraction evaluated on a sample of 228 out of 1500 of the orphan activities at the time).

These calls are probably part of the driving force that led to the recent efforts of the UniProt Consortium [3]. In parallel, [238] have set up the OrEnzA database that is dedicated to listing, for the most recent UniProt release, all the Orphan Enzymatic Activities (as defined across all organisms, be they prokaryotes, eukaryotes or archaea).

Here, I present a brief database study that David VALLENET and I carried out in order to establish the adoption status of the EC numbers used in UniProt and the MicroScope platform.

VI.B.1.a. EC numbers over the years

The number of orphan enzymes varies over the years due to several factors:

- new EC numbers are discovered
- EC numbers are modified or removed
- coding genes for EC numbers are discovered and are published directly
- bibliographic efforts associate coding genes to EC numbers in old publications
- annotations are modified, corrected or removed

Here, we established a snapshot of the evolution of EC number discovery and gene association, allowing us to represent how badly annotations lag behind the discovery of new activities.

Protocol: We wanted to evaluate, for each year since the beginning of molecular biology:

- the number of EC numbers discovered;
- the number of EC numbers successfully associated to a gene/protein sequence in a scientific publication.

The list of EC numbers and their year of creation are accessible in the ENZYME database. In order to evaluate the date of the first successful gene/protein assignment for a given EC number, we decided to keep the earliest publication giving an EC-sequence association, with the additional constraint of ignoring publications concerning more than 10 proteins (a heuristic for avoiding whole-genome publications that cannot be considered as biologically precise enough for a given enzymatic activity).

To obtain our primary data, we followed the following protocol:

- From the ENZYME database, extract the list of all current EC numbers (thus ignoring deleted numbers). We obtain an up-to-date list of EC numbers.
- From the UniProt KnowledgeBase, extract the list of all proteins associated to an EC number (be they from the SwissProt component or the TrEMBL component). We thus have an instantaneous “snapshot” of the annotation state of all EC numbers.
- From PubMed, extract the list and correspondences of all publications referencing proteins.

From this data we thus extracted, per year, the number of ECs discovered, and an estimation of the number of ECs assigned to a protein/gene sequence.

Results: Results are given in [Illustration VI.1: EC numbers over time]. Biochemists have actively discovered new chemical reactions catalysed by biological agents over the past century. Now-computerised scientific publications indicate the use of IUBMB-defined EC numbers since the mid-1930's. Other publications show that the discovery of the protein sequences of the responsible enzymes started shyly in the 1950's with the development of Sanger protein sequence sequencing. In the 1970's, the development of Sanger DNA sequencing greatly boosted the number of gene sequences associated to EC-bearing enzymes, while the development of Expression Cloning stimulated the discovery of novel enzymatic activities. Over the past decade, despite the completion of the genomic sequence of one, then thousands of organisms, both the discovery of new activities and of coding sequences have dropped considerably. More worryingly, this drop has not allowed sequence discovery to catch up with activity discovery. This explains the current 1,186 (27% of 4,150) of enzymatic activities that remain sequence orphans to this day.

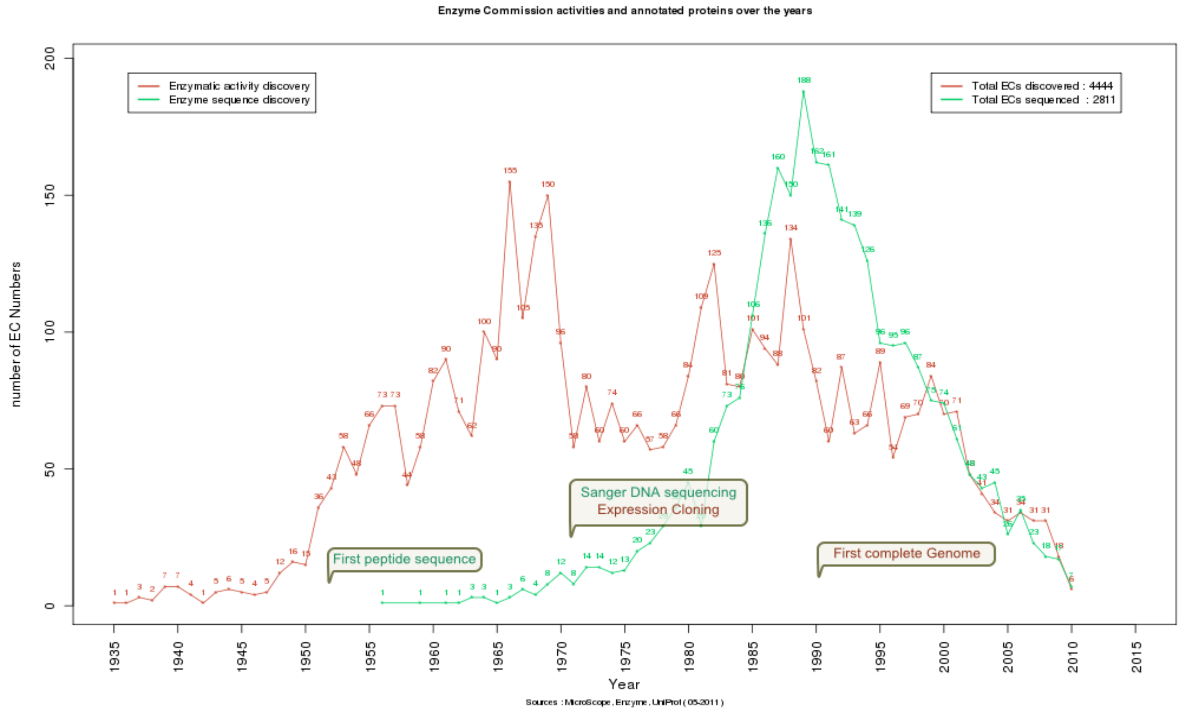


Illustration VI.1: EC numbers over time

Evolution of the number of declared ECs numbers and the number of those with at least one gene in UniProt.

These estimations are not fixed over time, however. Indeed, as already said, activity descriptions evolve, and bibliographic computerisation efforts allow improved access to results dating back to the start of the century. UniProt is typically undertaking such an effort [3]. The differences between the counts from 2007 and today are not graphically visible, yet the total number of orphans has fallen from 35% to 27% in the intervening time, despite an increase in EC number count (data not shown).

VI.B.1.b. Orphan activity counts

As a result of the previously represented history, a certain number of metabolic activities are considered globally orphan today. Establishing an up-to-date list of them is of prime importance for bioinformatics methods that focus on finding candidate genes for orphan enzymes.

Protocol: As before, the set of all EC numbers was extracted from ENZYME. All annotations were extracted from UniProt KnowledgeBase, KEGG and MicroScope's database. Finally, OrEnzA was interrogated.

Results: The following results were obtained in autumn 2010:

Database	Number of EC numbers	Percentage of total
ENZYME	4,150	100.00%
	Number of global orphan EC numbers	
UniProtKB	1,186	28.58%
- SwissProt	1,664	40.10%
- TrEMBL	1,440	34.70%
KEGG	1,938	46.70%
MicroScope	2,038	49.11%
OrEnzA	1,170	28.19%
All	1,132	27.28%

SwissProt has a higher level of global orphan EC numbers than TrEMBL, as the latter contains mostly (but not exclusively) bioinformatics predictions that have not been sufficiently verified for integration into the former.

MicroScope has the highest level of orphan EC numbers, however this is to be expected, as it is limited to prokaryote organisms only, while some EC numbers can be eukaryote-specific.

All of these resources, however, have a small number of non-orphan EC numbers that is specific to them (data not shown), likely due to differences in procedures and effort investment. This is underlined by the lower global orphan EC fraction when considering the union of all the resources (27.3%).

The OrEnzA resource should have exactly the same count of orphans as UniProtKB, as it is derived from it using the same protocol presented above; the observed difference is due to a difference in releases.

VI.C. Methods for finding candidate genes

As already stated, homology-based annotation methods are not capable of solving the orphan enzyme problem for lack of known sequences for the target activities, unlike context-based methods. A review of web-based tools that can be used manually by bioanalysts to find the most promising candidate genes for a target reaction (not necessarily orphan) can be found in [239]. Below, I review for my part several context-based methods specifically capable of proposing candidate genes for orphan enzymes.

VI.C.1.a. Annotation platforms

In section III.B.2, I pointed out how annotation platforms allow bioanalysts to act as human integrators of diverse sequence- and context-based information sources in order to brew up a plausible functional annotation of a target gene. A bioanalyst might thus be able to propose candidate genes for an orphan enzyme. One such success story can be found in [240]. However, bioinformatics aim (amongst other things) to develop increasingly automated methodologies, simplifying and speeding up manual analysis.

VI.C.1.b. Exploitation of STRING data

The STRING is not an annotation platform. Even though it does not provide functional predictions based on its associations, it has been manually exploited by its users to find candidate genes for missing enzymes in several showcases [241–243]. To my knowledge, however, no automated methods exploit STRING data specifically in order to find candidates for orphan enzymes. At best, [157] use STRING data with their “elementary mode”-based pathway reconstruction method in order to extract lists of candidate genes for reaction gaps in the purine and pyrimidine metabolic pathways.

VI.C.1.c. FRECs

[144] were the first to use metabolic context with genomic proximity to help locate sets of co-localised genes coding metabolically linked reactions, allowing gaps, which they called Functionally Related Enzyme Clusters (FRECs). Their method represents genomes as circular graphs with edges capturing gene co-localisation, ignoring transcription direction. Their KEGG-based metabolic network had reaction nodes linked by shared reactants, *i.e.* compounds that are products of one reaction and substrate of the other. Correspondences between the two are derived from existing KEGG EC number annotations.

Their algorithm is heuristic and searches for clusters of nodes in each graph that correspond via these annotations, allowing for gene and reaction gaps. It is similar to a hierarchical ascending clustering algorithm: starting from the set of all gene-reaction associations as initial clusters, the latter are aggregated progressively using a single linkage algorithm that processes a distance matrix derived from two separate distances, the one calculated on the gene graph, the other in the reaction graph. More specifically, if C_i and C_j are two clusters, $d_g(i,j)$ describing the distance in the gene graph between the clusters, and $d_r(i,j)$ the distance in the reaction graph, then the integrated distance $\delta(i,j)$ is a binary value that is only worth 1 when the minimum distance between all pairs of genes from C_i and C_j is less or equal to $1 + Gap_{gene}$, and when the minimum distance between all pairs of reactions from C_i and C_j is less or equal to $1 + Gap_{reaction}$.

Conservation of located FRECs across genomes was evaluated manually until the slightly later works of [143], which used the same graph comparison method along with another algorithm in order to locate a) conserved syntenies between pairs of genomes and then b) conserved syntenies across all genomes. This served as a basis for establishing KEGG's ortholog families (KEGG KOs).

These works were continued somewhat in [244], where the authors integrate several functional dependency indicators into a single indicator using a supervised learning approach. They teach a kernel canonical analysis algorithm [245] to predict protein-protein associations from kernel similarities calculated for gene cluster and phylogenetic profile data. Candidates for missing enzymes in known metabolic pathways are manually selected amongst the network neighbours of genes present in the given pathways, preferring genes already having annotations with partial EC number matches to the target reaction.

VI.C.1.d. ADOMETA

Various methods have been developed at Vitkup labs, the current show case being **ADOMETA**

(Adoption of Metabolic Genes for Orphan Activities), their orphan enzyme candidate gene finder that exploits genomic and expression context information [246,216]. Its general procedure is given in [Illustration VI.2: ADOMETA General procedure]. ADOMETA uses organism-specific metabolic reconstructions to define EC-number reaction-centred metabolic networks with degree-filtered compounds [Figure step 1]. Each reaction node is then “populated” with known enzyme-coding genes when available; unpopulated nodes are local orphan enzymes [Figure step 2]. The metabolic context of a target orphan is the set of all reactions connected to it by a path of maximum length k (k is typically set to 1, 2 or 3) [Figure step 3]. ADOMETA compiles several genomic context-based functional dependency measures: co-expression profile similarity, phylogenetic profile similarity, gene fusion/fission score, ordered gene clustering score, and protein interaction data. ADOMETA then builds an integrated score using two approaches, “direct likelihood-ratio” integration and integration using the ADABOOST [247] algorithm. These scores are then used to rank genes of unknown function for gene-less reaction nodes [Figure steps 4 and 5]. A candidate gene's functional dependency for a target gene-less reaction is a weighted combination of its functional dependencies with the genes populating the target reaction's neighbour nodes - any of the component dependency measures, as well as the two integrated measures, can be used. The ADOMETA web server gives access to all these scores and ranks candidate genes based on them. ADOMETA is the first resource, to my knowledge, that can effectively propose candidate genes for orphan enzymes using contextual information in an entirely automated approach.

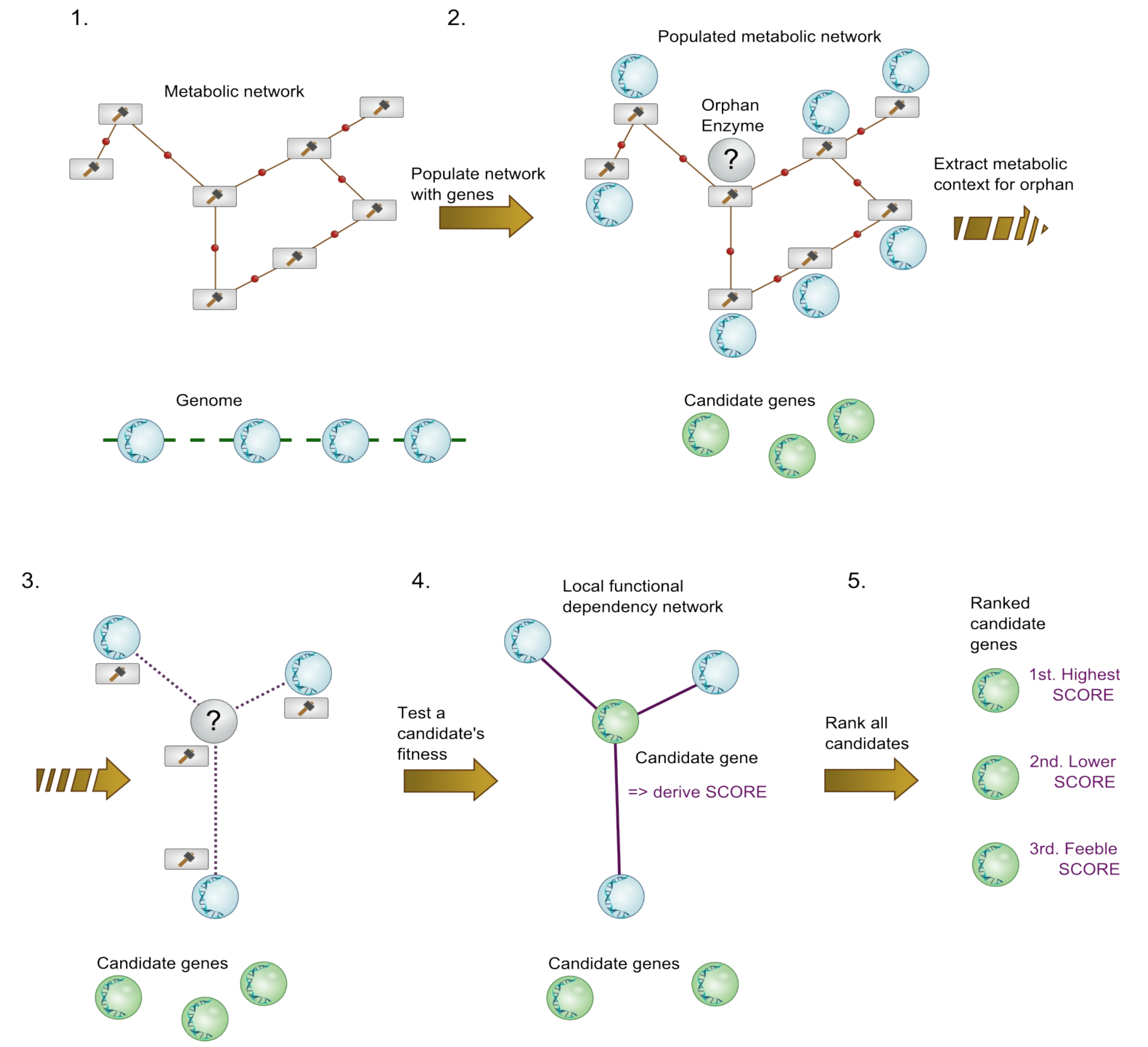


Illustration VI.2: ADOMETA General procedure

From left to right: 1. For a target organism, retrieve metabolic network and list of genes. 2. Populate reaction nodes with genes according to gene annotations. Unassigned genes form the candidate gene list. 3. Local functional dependency network (metabolic context) for a target orphan enzyme is the set of reactions connected to the orphan activity in the metabolic network. The local functional dependency network for a candidate gene is extracted. A “fitness” score is derived for the candidate gene based on its functional dependency scores with genes encoding reactions from the orphan reaction metabolic context. 4. All candidates are scored, then ranked.

VI.C.1.e. Pathway Hole Filler

A naïve Bayesian network-based approach called the **PathwayHoleFiller** (PHF) in [248] integrates genomic (BLAST data) and metabolic (reaction adjacency, pathway “directon”) information in order to improve homology-based functional transfer. However, PHF cannot work for global orphan enzymes. This shortcoming was addressed in [249,248], where the authors adapted their previous strategy to allow for context-based functional dependence. Firstly, they extracted protein-protein association data from Prolinks (gene neighbours, gene clusters, gene fusion, and phylogenetic profiles), in order to derive a binary functional dependence network. Secondly, they modified the strategy for proposing gene candidates for specific activities. In the case of orphan activities, not all genes from a genome were considered as candidates; rather, they were selected on the basis of three rules. Given a set of "seed" genes coding proteins catalysing other reaction steps of pathways containing the target orphan activity, candidate genes are necessarily 1) in the same transcription directon as at least one of the seed genes, or 2) in the immediate network neighbourhood (as defined by ProLinks functional associations rendered binary) of at least a seed gene, and 3) not encoders of proteins already known to participate in the *same* pathway (allowing existing annotations to be questioned). Finally, they extended their original Bayesian classifier with new nodes to take into account context-dependant information, though exactly *how* is not described in their paper [248]. These modifications allowed candidate genes for global orphan enzymes to be proposed, scored and ranked in all organisms with a PGDB. They validate their new approach by performing a 5- or 10-fold cross validation on a restricted list[10] of known reaction-coding genes from EcoCyc [207], as well as on a less well curated PGDB, CauloCyc.

VI.C.1.f. Miscellaneous

A couple of other methodologies might be worth mentioning for the more algorithmic-hungry readers. Yao *et al.* [250] use a k-nearest neighbour clustering approach in a feature space integrating several more traditional dependency measures (expression correlation, chromosomal distance, gene clusters, and paralogy) into one using a chosen kernel. Obtained clusters are assigned functional terms from KEGG maps, COGs, or MultiFun using a voting scheme allowing for confidence estimation. Chen *et al. [133]* use an algorithm based on path-walking through a gene co-localisation/sequence similarity-based graph reminiscent of SNAP to derive a novel measure of functional dependency. Missing enzymes in KEGG pathways are proposed and ranked on the basis of their functional dependency with genes already participating in those pathways.

The small number of accessible automated methods for finding candidate genes for orphan enzymes is one of the driving forces behind our intent to develop the "finding Candidate Genes for Orphan Enzymes" (CanOE) strategy, as presented in the following chapter.

[10] Namely removing multi-functional enzyme-coding genes.

VII. Development: The CanOE strategy

The CanOE ("finding Candidate Genes for Orphan Enzymes") strategy is described in the scientific article (and its associated supplementary material) that is included hereafter. Some points are discussed more in detail here, particularly concerning experiments and technical implementation that were a major part of my work and that are not put forward in the article. Finally, additional perspectives and considerations are discussed concerning the use of the strategy.

VII.A. Overview

As already stated, a large fraction (27% at latest estimate) of all EC numbers refer to sequence-orphan enzymatic activities. The use of context-based inference is a necessity, as usual homology-based annotation transfer methods cannot annotate genes with these activities. The aim of this work was to propose a new strategy, inspired by bioanalysts' manual *modus operandi*, that could use genomic and metabolic context information contained in MicroScope's database to propose candidate genes for global (or at least *prokaryote*) orphan activities. This is done using a graph-based algorithm that locates groups of co-localised genes coding enzymes catalysing groups of reactions that are "close" in the global metabolic network, specifically allowing for gaps. Potential associations are proposed between gene and reaction gaps, and prioritised using a score integrating results over all organisms by a family-based approach. Generated high-ranking functional hypotheses can then be verified experimentally, for example by the LGBM (laboratoire de génomique et de biochimie du métabolisme) and LCAB (laboratoire de clonage et de criblage des activités de bioconversion) teams at the Genoscope, which have a working history of elucidating orphan enzyme puzzles [240,251,252]. The strategy must be accessible to all MicroScope users, preferentially via dedicated web interfaces.

VII.B. Article

Author Contributions and Acknowledgements: *Alexander SMITH and David VALLENET designed the strategy. AATS implemented it, as well as the web interface. DV and Gregory SALVIGNOL adapted the web interface for its public deployment. DV and Eugeni BELDA directed the bioanalysis of the case study. Claudine MEDIGUE and Alain VIARI reviewed the manuscript. AV lead the development of the CCCPart algorithm and graciously granted us permission to use it in this application. Damien MORNICO and François LEFEVRE assisted with parsing MetaCyc data to the MicroScope database. Marcel SALANOUBAT and Alain PERRET are involved with the tentative biochemical validation of the CanOE case study presented in the article.*

VII.C. Used concepts and tools

The CanOE strategy uses few outside tools but each of particular importance.

VII.C.1. The CCCPart algorithm

The Common Connected Component Partitioner (**CCCPart** or simply C3P) was developed in 2005 by Frederic BOYER and co-workers [5], and of which I summarise the general *modus operandi* here (see [Illustration VII.1: The CCCPart algorithm], borrowed from [5]).

Consider two unweighted, undirected graphs G1 and G2, the nodes of which are linked (or not) by a set of binary relationships R. These two graphs can be summarized in a single graph, called the correspondence multigraph, the "multi-nodes" of which represent distinct pairs of nodes from G1 and G2 that are linked in R, and with two types of edges: "G1 edges" between multinodes of which the G1 vertices are linked by an edge in G1, "G2 edges" between multinodes of which the G2 vertices are linked by an edge in G2.

Connected Components are, intuitively, groups of nodes in a given graph that can all be reached from one another by walking the graph. Common Connected Components (CCCs) are the extension of this notion to multigraphs, and sets of vertices such that every vertex is reachable from each other vertex, through each type of edge taken separately.

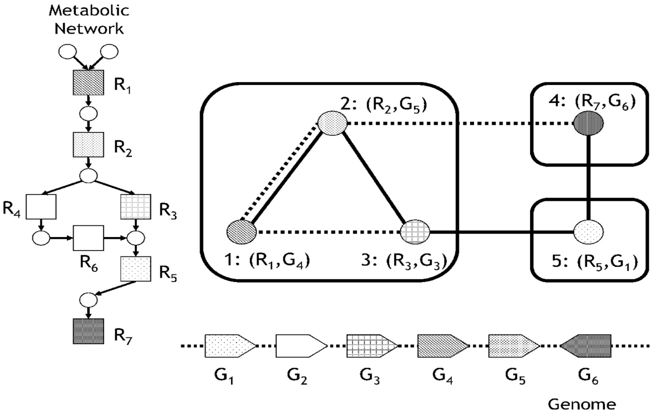


Illustration VII.1: The CCCPart algorithm

The metabolic network on the left and the gene graph on the bottom (with correspondences between the two shown by shared colours/hatches) are merged into a since graph, the MultiGraph, where each node corresponds to a (gene, reaction) pair, *i.e.* an annotation. Two types of edges can connect two multi-nodes, thos corresponding to metabolic adjacency in the metabolic graph, and those corresponding to gene adjacency in the gene graph. For example, (R2, G5) is connected to (R7, G6) by a gene edge because G5 and G6 are adjacent. CCCPart creates a partition of the multigraph, where each cluster (Common Connected Component, CCC) corresponds to a component that is connected for each type of edge (*e.g.* (R2, G5), (R1, G4) and (R3, G3) are in a same CCC because (R2, G5) is connected to (R1, G4) and (R3, G3) by reaction edges, and (R1, G4) is connected to (R2, G5) and (R3, G3) by gene edges).

The mathematical definition and algorithmic treatment of the problem of locating the set of CCCs between two input graphs are given in [5]. The algorithm is fast and deterministic, unlike previous similar approaches [144,253]. The algorithmic complexity is slightly worse than in [254,255], but [5] show that the time-limiting step is not CCC calculation, but multigraph construction. This has been greatly improved in following versions of CCCPart [146,147,256].

CCCPart and its more recent version have been used in several applications, such as locating syntons, or finding conserved modules between protein-protein interaction networks [6,256]. Here, we decided to apply it to locating units of metabolic function on prokaryote genomes, also known as “**metabolons**”, as suggested by the authors. This required some adaptations, especially the recovery of metabolon gene and reactions gaps from the CCCs returned by the CCCPart algorithm. The most important details of this development are described in the paper's Supplementary Material.

VII.C.2. OrthoMCL

The process of building gene/protein families for the multi-organism metabolon integration step was done by a home-made program heavily inspired from the **OrthoMCL** software [68], a well-known program that had already proven itself for protein clustering on sequence similarity [257].

The reasons we did not use the original program for CanOE were: a) the available version of OrthoMCL at the time (version 2, published in February 2008) did not allow us to use protein similarity data already stored in the MicroScope database, requiring complete recalculation of similarities, which was not tractable; b) the available version of OrthoMCL at the time was implemented in Perl, a scripting language none of the LABGeM team were comfortable with; and c) studies by myself and Damien MORNICO seemed to suggest that the then-available version of OrthoMCL actually had bugs that invalidated some of its results. OrthoMCL has now been completely overhauled by its authors (now available: version 5 published in March 2011), but we will continue to use our own version until the MicroScope policy on explicitly integrating gene families into its annotation pipeline has been determined.

The principle of OrthoMCL is relatively straightforward: using all-against-all protein sequence similarity results from BLAST, it builds a n*n similarity matrix (where n is the number of proteins) summarising the similarities by scores. These scores are normalised following a procedure that deprecates similarities between phylogenetically close organisms, as well as taking into account inferred ortholog/paralog relationships between proteins. This matrix is then submitted to MCL, the Markov Clustering Algorithm devised by Stjin van Dongen [67], a now well-known and proven weighted graph-clustering algorithm that has since been applied to many different problems [258]

The MCL algorithm is based on the notion of random walks in a graph: two nodes should cluster together if the probability of randomly walking from one to he other is high. Simply put, the MCL algorithm iteratively updates positive edge weights (possibly creating or removing edges), favouring those that correspond to high random walk probabilities during an “expansion” step, and down-favouring those that have lower probabilities in an “inflation” step. The process (almost always) converges, and final clusters can (almost always) be identified as distinct connected components in the final graph. Many indicators are available at program termination, allowing a user to manually assess how well the clustering performed (such as clustering agreements, sum of edge weights preserved/cut, number of singleton nodes, and many many more).

VII.C.3. Metabolic Data used in CanOE

In the CanOE article, the only metabolic data used is that from the MetaCyc database; however, in other experiments, data from the KEGG database were used, and others could be. Indeed, CanOE's implementation surrounding CCCPart was designed to allow a certain degree of freedom regarding the source of metabolic data, *i.e.* metabolic database specificities are NOT hard-coded into the program. Rather, metabolic database data is parsed into a CanOE-specific standardised MicroScope database prior to running, and is retrieved by CanOE using generic Structured Query Language (SQL[11]) scripts.

In the article, we presented results calculated off a metabolic network built from MetaCyc data [259]. However, during conception of the CanOE strategy, several different protocols were tested out for creating different metabolic networks, referred to as “metabolic schemas”. These other schemas will briefly be described here, as well as the reasons behind our final choice of publishing only the MetaCyc schema (though use of KEGG as a metabolic data source is suggested in the supplementary material).

The first metabolic networks I tested were based on KEGG data [260] stored locally in MicroScope. Reactions were KEGG reactions (removing multi-step reactions for which each step had a corresponding specific reaction). Gene-reaction associations were created based on gene-EC associations from MicroScope and EC-KEGG reaction correspondences. One EC could correspond to many KEGG reactions and vice versa, posing some spurious multiplicity problems, which were the primary impetus for the development of the metabolon gap-filtering step mentioned previously.

VII.C.3.a. “Main KEGG Reaction” Schema

The “main KEGG reaction” schema was a first attempt at translating data from KEGG's LIGAND database into a global network of reactions connected by edges when biologically relevant “main” compounds were shared (*i.e.* an important product of one reaction was an important substrate of the other). At the time, I had not found any way to extract the notion of “main compound” from the KEGG data, and resorted to an *ad hoc* approach already used in previous works (such as [261,216,262] and many others): some ubiquitous compounds were not used to generate edges (such as water, oxygen, carbon dioxide, protons...), and others were eliminated based on the high number of edges their use would generate (compounds from KEGG's specialist GLYCAN database were not used either, though most had corresponding compounds in LIGAND). Also, KEGG reactions that were known to be multi-step reactions (of which each known step already had a corresponding KEGG reaction) were ignored. Finally, given that the assignment of KEGG reactions to MicroScope genes was based on EC number correspondences, I was forced to ignore KEGG reactions that corresponded to more than 20 distinct EC numbers in order to avoid creating simply

[11] For some additional detail on SQL, see part VII.D.1

unbelievable metabolons.

Metabolons obtained with this schema generally had high gene-reaction multiplicity due to the EC-KEGG reaction conversion and showed many cases of biologically irrelevant reaction associations due to reaction-reaction edges being created from non-main compounds that the previous heuristic failed to recognise as such (*e.g.* acetyl-CoA).

KEGG Reaction-based networks can be hard to interpret without representing Compounds as nodes.

VII.C.3.b. “Main KEGG RPAIR” Schema

The “main KEGG RPAIR” schema was a later attempt to achieving a more relevant metabolic network (with less spurious edges) by exploiting not reaction descriptions, but RPAIR knowledge [263]. **RPAIRs** (reaction compound pairs) are defined for each metabolic reaction as pairs of compounds between which chemical functional groups were transferred during the reaction, and are instrumental at computationally following atom trajectories through a metabolic pathway [264]. They are created by a chemical graph alignment algorithm and are manually curated to ensure biological relevance. RPAIRs can be classified into different types, the most remarkable of which are “main” (*i.e.* determines exchange of functional groups between “main” compounds), “trans” (*i.e.* determines transfer of functional groups from a secondary metabolite to a main one) and “leave” (*i.e.* for a functional group that becomes an isolated compound, as in decarboxylation). RPAIRs have been used as a proxy for main compound definition in order refine KEGG metabolic knowledge in previous works [223,217]. They are, however, even more difficult to interpret in a network than Reactions. As to the lack of Compound representation, there is the added complexity that several RPAIRs can belong to a same reaction.

In this first RPAIR-based metabolic schema, the metabolic network nodes were RPAIRs rather than Reactions. RPAIRs were connected via edges corresponding to shared compounds. Only RPAIRs of type “main” were used, ensuring the use of “main” compounds in connecting the RPAIRs. As with the main KEGG reaction schema, I filtered out RPAIRs belonging to mutli-step reactions, as well as those corresponding to too many EC numbers to be usable.

Despite the use of “main” RPAIRs only, the resulting network was rather dense, even after filtering. This is probably due to the high number of RPAIRs that exist, and to the fact that many reactions have several RPAIRs, of which one or more can be “main”.

VII.C.3.c. “Main KEGG RPAIR-to-Reaction” Schema

In order to address the difficult interpretation of the RPAIR-based schema, I decided to try to use a Reaction-centric network, with edges defined on the basis of RPAIRs. Put simply, the “main” RPAIR-based network presented in the previous schema was transformed, replacing groups of RPAIRs by their corresponding Reactions, while conserving connectivity. The resulting network was sparser than the KEGG RPAIR one, but denser than the KEGG Reaction one. Furthermore, it was difficult to establish if this schema actually outperformed the previous two or not.

VII.C.3.d. “KEGG MAP” Schema

KEGG is a resource in constant development. During my thesis, KEGG released XML-formatted files for their KEGG Maps, each of which describes large sets of reactions belonging to a high-level metabolic process, such as glycolysis (ID: map00010), purine metabolism (ID: map00230) or the sum of all metabolic pathways (ID: map01100). Of particular interest to us here is the fact that reactions are linked in these maps by KEGG-defined main compounds. Unfortunately, due to 1) the large sizes of KEGG maps (which include several of what biochemists would usually consider as metabolic pathways), 2) the existence of errors in the XML data, and 3) the possibility for a same compound to appear in several disconnected locations on the same map, we were unsatisfied with this schema. The generated metabolons still contained many spurious reaction-to-reaction edges or missed metabolons altogether, and we were forced to abandon the use of this otherwise promising metabolic schema.

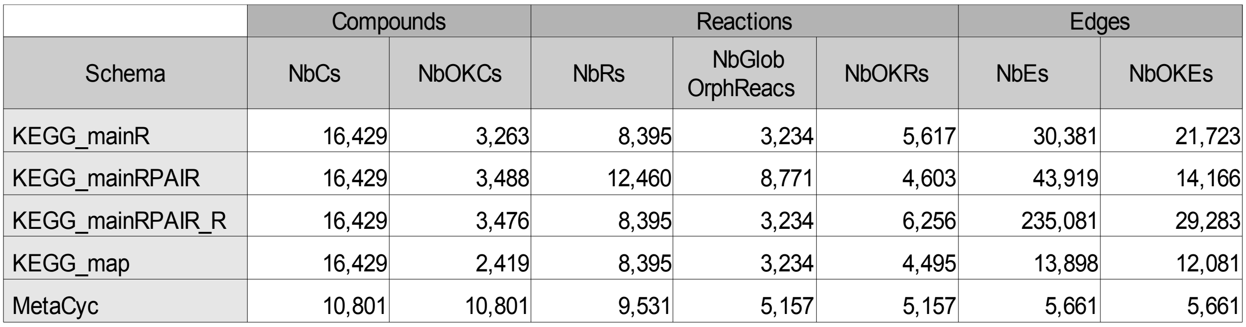
VII.C.3.e. KEGG afterword

KEGG was and still is an important source of computationally-formatted metabolic data that is widely used in the field of bioinformatics, even though the resource itself was never designed nor funded as a public database [http://www.genome.jp/kegg/docs/plea.html]. Unfortunately, due to concurrent circumstantial pressures, Kanehisa laboratories are no longer able to fund the maintenance and development of KEGG beyond that of KEGG MEDICUS, a medically-orientated subset database. KEGG data shall remain web-accessible, but will only be available for download for paying subscribers, though the fee shall be cheaper for purely academic clients.

VII.C.3.f. “MetaCyc” Schema

After the disappointing results with KEGG-based schemas, and thanks to the continued development and integration of MetaCyc data into MicroScope, we decided to turn to this metabolic data resource instead. In the "MetaCyc" metabolic schema, two reactions are linked by an edge when the product of one is a substrate of the other. This information is readily available in the MetaCyc data system, and has been parsed to the MicroScope "MicroCyc" database. To avoid the high connectivity problems that are common when building metabolic networks (as was the case in the KEGG-based schemas), we limited such shared compounds to "main" compounds. In the original MetaCyc data, "main" compounds can be extracted relatively easily for a pair of reactions belonging to a same metabolic pathway. However, in order to link reactions between pathways, we relied on the pathway-level inter-pathway connections also available in the MetaCyc data. These connections allow the reactions of one pathway to be linked to a reaction of another pathway, and we considered "main" any product of one that was substrate of the other. Note that we only considered "reciprocal" inter-pathway links, *i.e.* for reaction R1 in pathway P1, and reaction R2 in pathway P2, it was necessary for P1 to link R1 to P2, and for P2 to link R2 to P1, for R1 and R2 to be connected. We enforced this condition as there are still many "imprecisions" in the MetaCyc data that would otherwise lead to incorrect pathway connections.

Some numbers describing the presented metabolic network schemas are given in the table below:



	Compounds		Reactions			Edges	
Schema	NbCs	NbOKCs	NbRs	NbGlob OrphReacs	NbOKRs	NbEs	NbOKEs
KEGG_mainR	16,429	3,263	8,395	3,234	5,617	30,381	21,723
KEGG_mainRPAIR	16,429	3,488	12,460	8,771	4,603	43,919	14,166
KEGG_mainRPAIR_R	16,429	3,476	8,395	3,234	6,256	235,081	29,283
KEGG_map	16,429	2,419	8,395	3,234	4,495	13,898	12,081
MetaCyc	10,801	10,801	9,531	5,157	5,157	5,661	5,661

Illustration VII.2: Summary Description of Metabolic Network Schemas

Each row corresponds to a metabolic network schema, from top to bottom: KEGG_mainR (main KEGG reaction schema), KEGG_mainRPAIR (main KEGG RPAIR schema), KEGG_mainRPAIR_R (main KEGG RPAIR-to-Reaction schema), KEGG_map (KEGG map schema) and MetaCyc (MetaCyc schema). Columns from left to right: NbCs (number of compounds), NbOKCs (number of compounds having passed pre-processing), NbRs (number of reactions), NbGlobOrphReacs (number of global orphan reactions), NbOKRs (number of reactions having passed pre-processing), NbEs (number of edges), NbOKEs (number of edges having passed pre-processing).

The MetaCyc schema has the highest number of compounds kept after the pre-processing filters as, indeed, no filters based on compound usage or reaction degrees are used in it. However, it also has the sparsest network in terms of edges. Though this might reduce the number of alternate paths between any two reactions, we are assured that - thanks to the network construction protocol - *all* of these edges are of biological relevance, unlike in the KEGG-based schemas.

VII.D. MicroScope & CanOE data models

The CanOE strategy uses data and stores its results in MicroScope's PkGDB database system. It is thus necessary to present the latter in order to detail CanOE's implementation.

VII.D.1. MicroScope PkGDB

PkGDB is MicroScope's relational database system. Relational databases are a particular type of database, wherein different types of data are modelled with various n-ary relationships between them [265]. They are particularly useful in storing large amounts of structured data. Actually accessing the data requires the use of a querying language. SQL ("Structured Query Language") is the most widely used scripting language for constructing and querying relational databases. MySQL is an open-source relational database management system including both a database server and an SQL querying interface. PkGDB is a MySQL database system, and working with it requires knowledge of the way its data is modelled. Here, I shall present the databases and tables within PkGDB that are most relevant to my work and to the comprehension of the MicroScope platform as a whole.

VII.D.1.a. Primary Data

As explained in [6,4], MicroScope collects and compiles primary data from many different sources in order to propose the most complete vision possible of genomic sequence annotation. Such sources include UniProt, KEGG, MetaCyc, ENZYME, COG, InterPro, and others. Data from these sources are generally organised into specific PkGDB databases, and can be queried in conjunction with MicroScope-specific data.

VII.D.1.b. Core MicroScope data

The core MicroScope-specific data is contained in the "pkgdb" database. MicroScope is a platform dedicated to annotating prokaryote genomes (see chapter IV for details on annotation). The first tables of interest are thus those that describe the genomes contained within. Table "Organism" contains one row per prokaryote organism whose genome is available in MicroScope. Rows in table "Replicon" describe, for each organism, the various DNA molecules that compose the organism's genome (chromosomes, plasmids, megaplasmids...). Finally, table "Sequence" describes the available DNA sequences available for the previously described replicons. Indeed, organisms can be resequenced (or more often re-assembled and finished with more recent methods), rendering previous sequences obsolete. Additionally, this table tracks the public or private status of some of MicroScope's organism sequences (as some projects are submitted for private annotation), as well as the advancement of the sequence through MicroScope's automated annotation pipeline (syntactic or functional). Now that individual sequences have been defined, I can expose how individual “genes” are modelled in the database.

The atomic unit of the PkGDB database system is the Genomic Object (GO). Current implementation defines a Genomic Object as a short stretch of genome (generally a CDS, a falsely-predicted CDS, or ribosomal/transfer RNAs) that has been annotated (either by the automatic pipeline or by an expert bioanalyst). With each new annotation of a given stretch, a new Genomic Object is created, with a historical reference to the first Genomic Object for that given stretch (if any), ensuring that annotation history can be recovered if required. Annotation information contained in this table include Genomic Object type (CDS, fCDS[12], rRNA, tRNA...), sequence frame (-3 to +3), start and stop positions, annotation status (automatic annotation finished, artefact, curated, in progress...), a reference label, gene names & synonyms if any, a description of the GO product, comments, any EC numbers or MetaCyc reactions in the case of a metabolic gene, and whether the GO is obsolete or not, amongst other things. Given the number of organisms, sequences, and the archiving of annotation history, this table is very large (32 columns, >7.6.10^6 rows, 4.7 Gigabytes on the 30th of August 2011).

[12] fCDS: “fragment of coding sequence”, typically a CDS that has been broken during evolution by one or several nucleotide mutations (*i.e.* a pseudogene).

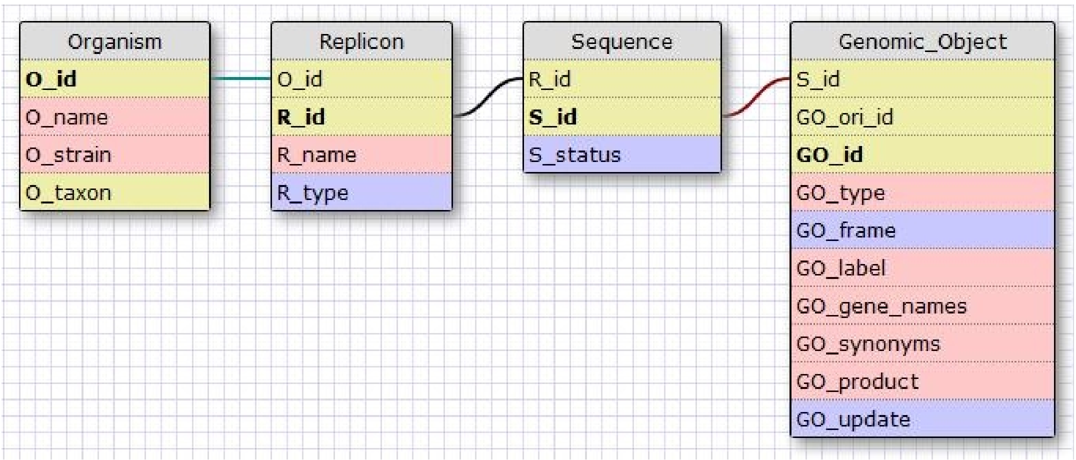


Illustration VII.3: The PkGDB Genomic Object Model

One Genomic Object (identified by a GO_id) corresponds to one annotation of a stretch of nucleotides. Re-annotating a previous stretch generates a new Genomic Object; previous versions are referenced by the 'GO_ori_id' field which points to the GO_id of the first Genomic Object having been created on the given stretch of DNA. Genomic Objects belong to a given Replicon sequence (one replicon can have several Sequences, but only a single one that is up-to-date). An Organism can contain several Replicons (chromosomes, plasmids, *etc*).

VII.D.1.c. Predicted annotations

The predicted annotations established by MicroScope's automatic pipeline are stored in separate tables, some of which in the “pkgdb” database. One notable exception to this is the tables containing BLAST and synteny data for all-*versus*-all protein pairs, which are stored in the “GO_CPD” database. Obviously, all-*versus*-all protein comparisons require a lot of room (the GO_CPD PkGDB database currently takes up 4 terabytes, though it also contains other data than sequence similarities, such as phylogenetic profile data), and querying a single table containing all data would be computationally infeasible. To address this problem, the data was split into multiple tables, basically one table per genomic Sequence containing all GO BLAST results against GOs from all other genomic Sequences, and tracking which BLAST hits are part of a synteny or not. Fusion/fission data is also stored in this manner.

VII.D.2. CanOE database

CanOE exploits gene and reaction data as extracted from the MicroScope database. However, during its initial conception phases, it appeared that running CanOE would be computationally intensive, thus proscribing regular atomic updates for each new annotation made using the platform.

It was thus necessary to insure that CanOE would be run at regular, not-too-frequent intervals, and that the underlying data did not change. The CanOE database was designed to host a working copy of relevant MicroScope data, formatted for quick access, without losing references to "active" MicroScope data. Here, I shall present an ideal database structure for CanOE as I see it, though its implementation is not yet optimised, though it should be during the development of the deliverable version of CanOE. I might thus make references to the improvements to PkGDB and the CanOE data models that I propose in section VII.D.3.

VII.D.2.a. Primary data: Genes and gene data

The CanOE database contains a copy of the latest data from the MicroScope "pkgdb" database, a sort of snapshot taken at a given point in time that is recorded. The Organism, Replicon and Sequence tables are present, as in "pkgdb", though only current public & private sequences are allowed, leading to a 1-to-1 relationship between replicon and sequence ids. The Genomic Object table is split in such a way as to optimise some requests. First, no annotation history is required, thus eliminating the need for a historical table. Second, a correspondence table tracking sequence and genomic object ids is stored, called Genomic_Object_ID (akin to the Genomic Object table proposed in my new data model in later section VII.D.3). Genomic_Object_Data contains all GO-specific data that is relevant to CanOE (akin to the Genomic Object Annotation proposed in my new data model).

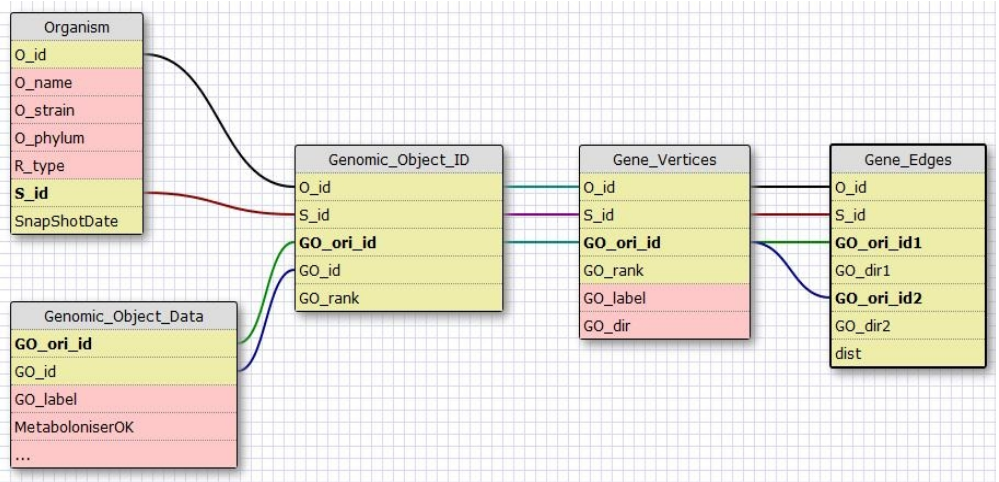


Illustration VII.4: Schema of the CanOE Genomic Object Data model

The Genomic Object data model, in CanOE, is composed of Organism objects, Genomic Objects, corresponding Gene Vertices, the Gene-to-gene Edges. Not all foreign key relationships are shown for clarity.

VII.D.2.b. Primary data: Metabolic Network Reactions

Metabolic data is saved using a different approach. The database was designed to store data parsed from MicroScope metabolic databases into a CanOE-specific data model. In CanOE metabolons, reactions are represented by vertices and edges correspond to main compound sharing. The model was designed so that metabolic reactions could be extracted from any type of metabolic database (*e.g.* MetaCyc or KEGG). I called a “Metabolic Schema” the parsed data for a given metabolic database extracted following a given protocol. Table MNW_Schemas (MNW is the acronym of Metabolic NetWork, and is a prefix for all CanOE metabolic data tables & IDs) describes the various available Metabolic NetWork schemas; table MNW_Reactions associates a distinct ID to any kind of metabolic reaction.

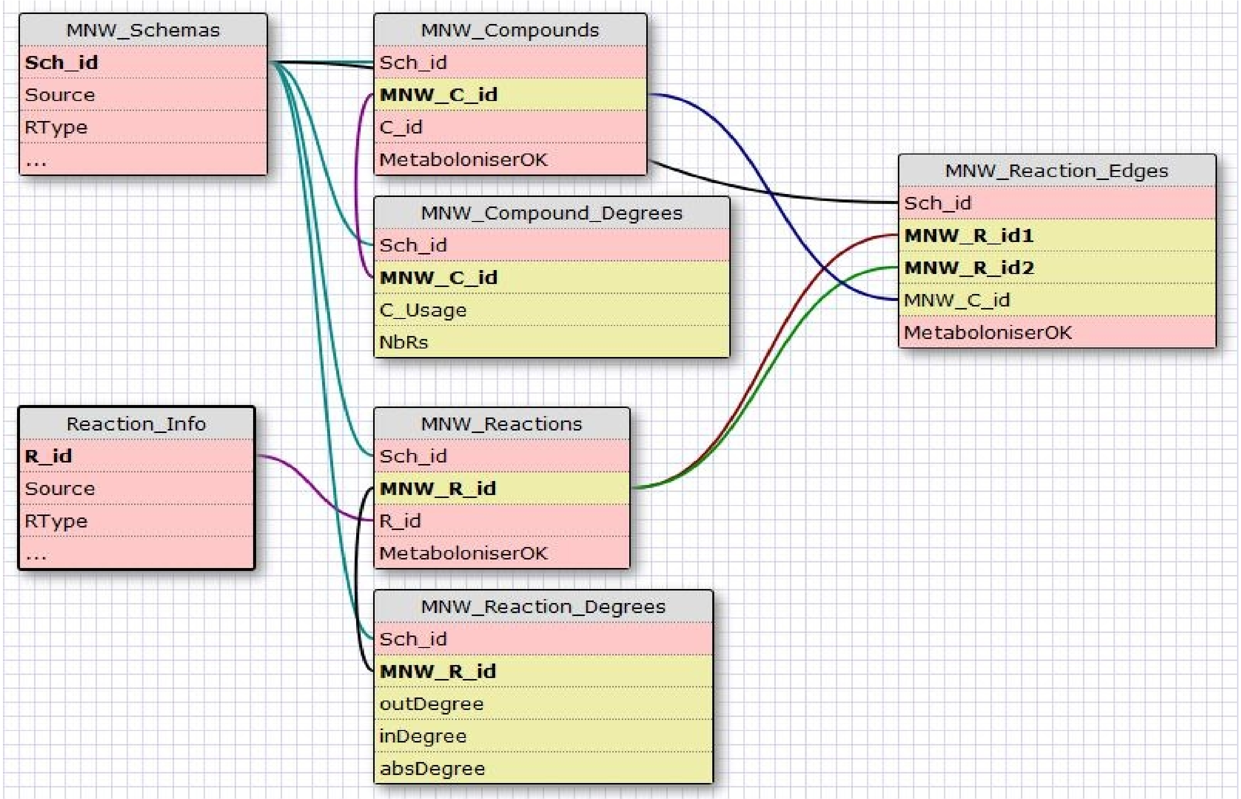


Illustration VII.5: Schema of the CanOE Metabolic Network Model

The CanOE Metabolic Network Model is composed of metabolic Schemas, Compounds, Reactions and Reaction-to-reaction Edges. The Reaction_Info table contains data from the source metabolic database of each Reaction. Compound and Reaction degrees are stored in separate tables for the filtering phase. Not all foreign key relationships are shown for clarity.

VII.D.2.c. Primary data: Gene-reaction associations

Thirdly, correspondences between GOs and metabolic reactions are stored in separate tables. Two types of association are kept. Metabolic data was initially stored in MicroScope using EC numbers (though since MetaCyc reactions have been added), so a GO_EC_CPD table was thus created in CanOE in order to store a snapshot of all (GO,EC) annotations for all GOs. Then, schema-specific SQL scripts (dependant or not on the GO_EC_CPD table) are executed in order to fill a GO_MNWR_CPD table that tracks correspondences between GOs and schema-specific reaction identifiers.

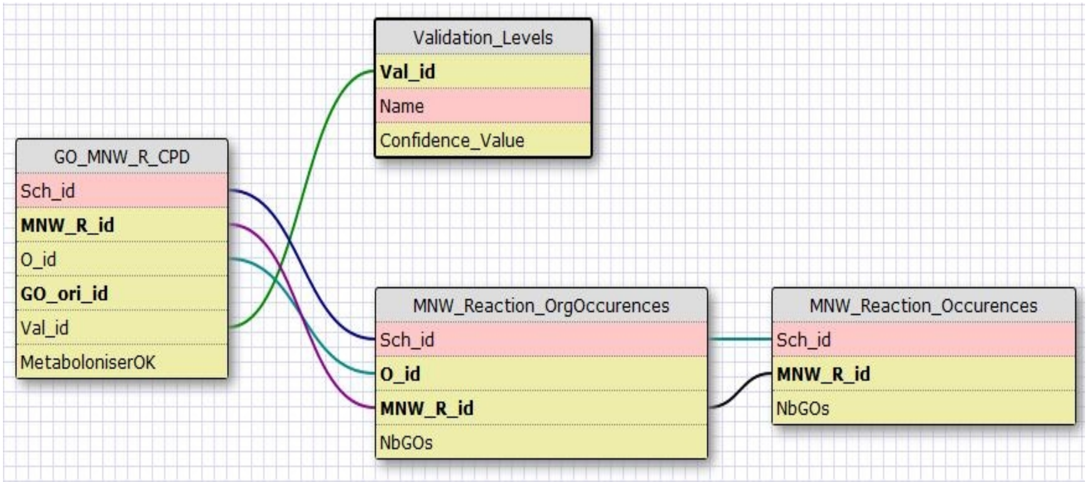


Illustration VII.6: Schema of the CanOE Primary association tables

The primary association tables describe the initial knowledge about gene-reaction associations given to CanOE. Reactions occurences are accounted for per organism and globally, in order to determine per-organism and MicroScope-wide orphan reactions.

VII.D.2.d. Primary data: gene and reaction graphs

In order to simplify and speed up the execution of the Metaboloniser algorithm, gene and reaction graphs are prepared beforehand and stored in the database (see [Illustration VII.4: Schema of the CanOE Genomic Object Data model]). The Gene_Vertices and Gene_Edges tables describe the gene graph, with indexed references to organism and sequence IDs to speed up queries. The MNW_Reaction_Edges table contains reaction graph edges, and the already-existing MNW_Reactions tables describes the reaction graph vertices (see [Illustration VII.5: Schema of the CanOE Metabolic Network Model]).

VII.D.2.e. Secondary data: Metabolons

A metabolon is a graph containing two types of nodes (gene, reaction) and three types of edges (gene-reaction associations, gene-gene adjacency, reaction-reaction compound sharing). Furthermore, the gene-reaction associations are of several types: Known annotation, Potential association, and Inferred annotation. It was thus necessary to come up with a data model that might save all this information the most efficiently as possible.

The first obvious step was a table listing all found metabolons in all organisms, containing several indexes for easy and rapid access. The Metabolon_List table compiles references for the organism and the sequence, as well as containing an automatically-incremented primary ID for each metabolon. This table can then be joined to the other tables describing the contents of each metabolon.

The gene-gene and reaction-reaction edges are already stored as primary data and do not need to be saved again. The “most defining” objects in a metabolon are, due to the way CCCPart works, the Known gene-reaction associations it contains, though these are also primary data, and are not an intuitive way of dealing with metabolons. I thought it simpler to save, for each metabolon, the list of vertices it contains; edges would then be recoverable using these and the primary data. Tables Metabolon_Vertices_GOs and Metabolon_Vertices_Reactions store, for each metabolon ID, the GO_ori_id or MNW_R_id for its genes or reactions, respectively. It also stores whether the vertex is primary (*i.e.* participates in a Known association) or secondary (gap). However, generated potential gene-reaction associations also need to be stored, and MinPathLengths need to be calculated for them and for the Known associations. Thus, I created the Metabolon_Assocs table that stores for each metabolon ID, the list of (GO_ori_id, MNW_R_id) pairs that it contains, with the MPL and its type (“Known”, “Potential”, or “Imaginary”, a type added for benchmarking purposes[13]).

[13] Reminder: Inferred associations are created from Potential associations later on in the CanOE pipeline.

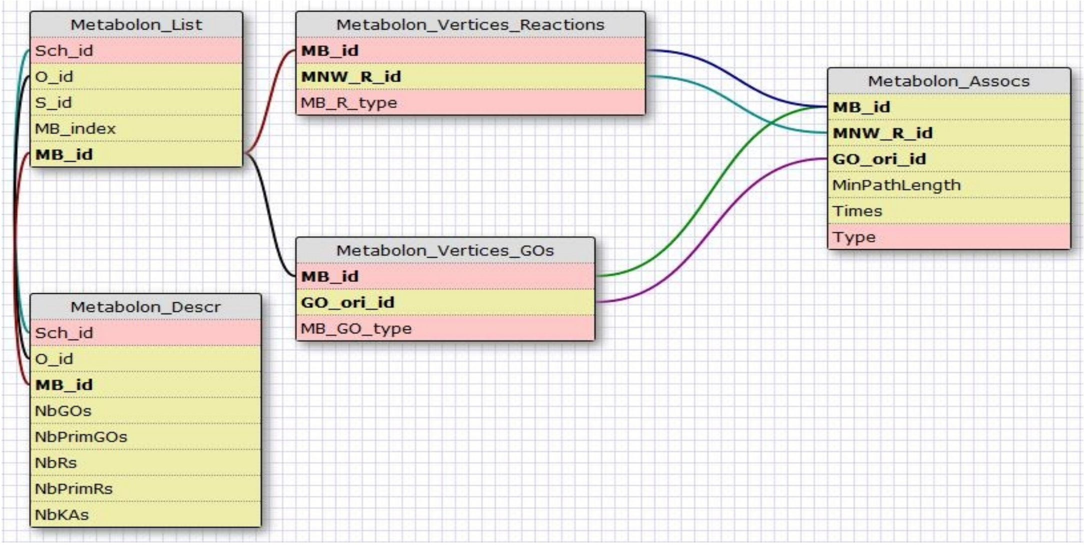


Illustration VII.7: Schema of the CanOE Metabolon data model

Metabolons are described by their component vertices (Genomic Objects and Reactions), and by the gene-reaction edges they contain, be they Known or Potential (pre-integration) ones (gene-gene and reaction-reaction edges can be recovered from the primary data). A metabolon summary table is also available.

VII.D.2.f. Tertiary data: Gene families

For the next steps of the CanOE strategy, the gene families established by our OrthoMCL-like protocol are required (see part VII.C.2). The results of the gene-family building algorithm are stored in the database in the MCL_clusters table, in the form of (GO_ori_id , CL_id) associations. Additional family information is stored in the MCL_CL_Data table, such as cluster size and family metabolic status as determined by the Gene Ontology-based protocol (described in the article's Supplementary Material).

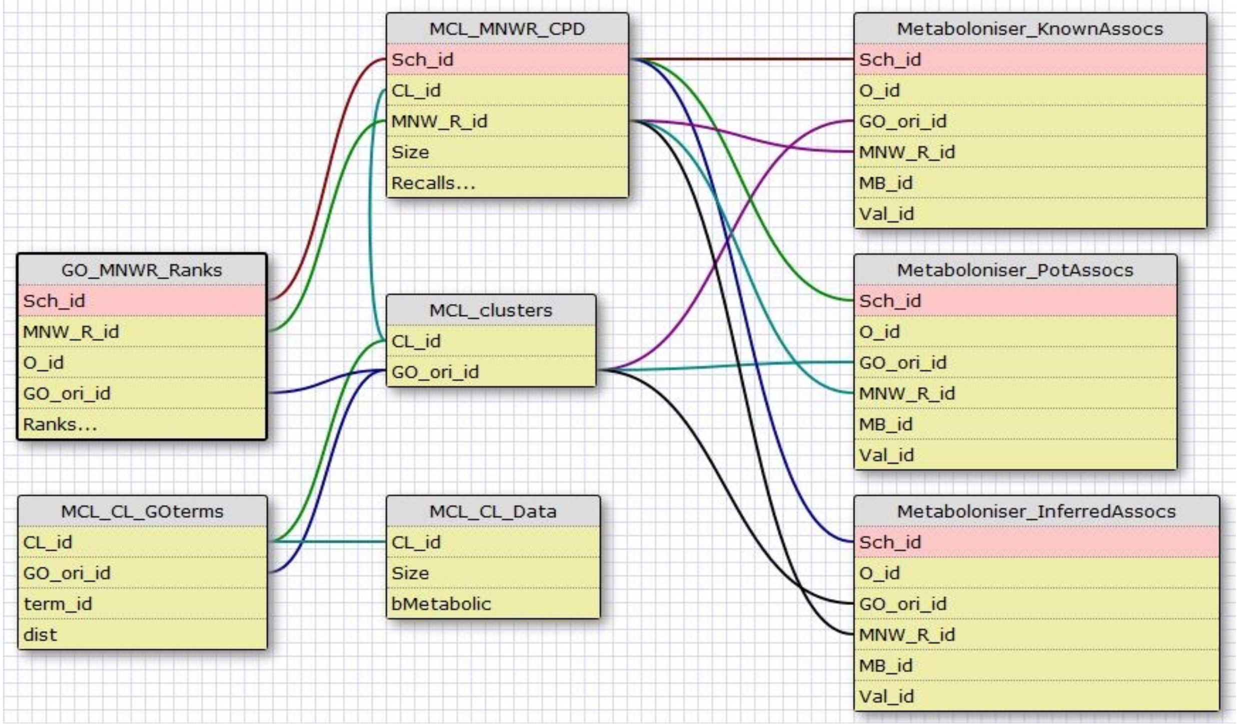


Illustration VII.8: Schema of the CanOE Metabolon integration data

The central table for the integration of CanOE results across organisms is the MCL_clusters table, describing gene assignments to gene families. Family-wide data (notably which families are metabolic or not) is stored in MCL_CL_Data. Metabolon-based gene-reaction associations are parsed into the Metaboloniser_KnownAssocs, Metaboloniser_PotAssocs, and Metaboloniser_InferredAssocs tables, and are integrated into family-reaction associations (with scores) described in table MCL_MNWR_CPD. The scores are then used to rank genes in each organism for each reaction, ranks that are stored in GO_MNWR_Ranks.

VII.D.2.g. Tertiary data: Known, Potential and Inferred associations

The previously-saved metabolon gene-reaction associations are processed during the CanOE pipeline into distinct tables according to their type. Known associations are saved in table Metaboloniser_KnownAssocs as they are. As described in the article, Potential associations (from table Metabolon_Assocs) are then processed into Potential associations (in table Metaboloniser_PotAssocs) and into Inferred associations (in table Metaboloniser_InferredAssocs), taking care to delete Potential associations corresponding to non-metabolic gene families or to genes/reactions that are no longer gaps in the inference step. See [Illustration VII.8: Schema of the CanOE Metabolon integration data].

VII.D.2.h. Tertiary data: Family-Reaction Association scores and ranks

The previous family and gene-reaction association data is compiled into family-reaction association data and transformed into the R=>F and F=>R scores using the formulae described in the article. These scores are stored in table MCL_MNWR_CPD. Then, for each organism and each reaction, the list of candidate genes for that reaction within that organism are ranked according to each score, and these ranks are stored in table GO_MNWR_Ranks. See [Illustration VII.8: Schema of the CanOE Metabolon integration data].

Taken altogether, the data stored in the previously described tables can describe in full detail all the objects and results obtained in the CanOE strategy, and it is queried in the CanOE web interface presented in section VII.F.2. However, the data models found within PkGDB, including those of CanOE, do not always conform to the ideal versions presented here. Even the latter could bear improving in some points. Having given the matter some thought, I have decided to lay my ideas down in writing here.

VII.D.3. Proposed PkGDB & CANOEDB improvements

MicroScope has grown over the years from a tool dedicated to the annotation of a single genome to a full-blown comparative genomics prokaryote annotation platform containing over 1,400 genomes. The production, database and visualisation systems (hardware and software) behind it have also evolved and are still perfectly capable of handling the requirements of over 6,000 annotations per month[14]. However, given the exponential increase in sequenced genomes, the MicroScope platform is expected to hit its limit in the years to come. Efforts are being planned to address this problem, such as the recruitment of a database specialist.

Indeed, one of the most criticised points on which improvements will have to focus is the design of the database. Though I am in no way a database specialist, I have decided to transcribe here my thoughts on this problem, in the hope that they might speed up future work. In any case, they shall allow the reader to understand the discrepancies found between the PkGDB database system, the implementations and the ideals I discuss in this manuscript concerning my thesis projects.

First of all, some **renaming** efforts are required. Foremost, the whole database system should have a distinct name from its child databases. Furthermore, the "pkgdb" database should have both a development version and a production version. The development version would obviously be used by the development team for tests, limiting the risk of messing the production database up, and not blocking SQL access to it for users. Currently, there is only a one database which is both for development and production.

As pointed out previously, my database designs relevant to my projects are somewhat different to those found in the PkGDB database system; indeed, I propose an alternative data model for Genomic Objects. The objective of the new model is multiple: it would be more efficient, especially space-wise, and also more intuitive, than the current one. The rationale behind it is basically to a) split Genomic Objects into Genomic Objects (GOs) and Genomic Object Annotations (GOAs), and to b) keep historical annotations in a different table than current annotations.

Genomic Objects should refer in a fixed way to specific DNA stretches on a genome sequence (though start and stop positions can be modified, and objects can be declared obsolete). Genomic Object Annotations would then be associated to these objects. Indeed, a GO does not evolve much over time, though multiple GOAs can be created for a single GO. Since only one GOA is the most recent, then this one could be kept for rapid access in a GOA table; historical GOAs could be transferred to a historical GOA table for future reference (especially since consulting the annotation history in MicroScope requires a specific action from the user). This would allow the main GOA table to be much smaller (a lot less rows, and no "GO_update" index column) without losing data accessibility[15].

[14] Average over year 2010.

[15] The same strategy could be applied to the Sequence table, that would then become obsolete: the Replicon table itself could contain data for the latest Sequence, and past Sequences could be kept in an OldSequence table. This idea is not presented in the figure on the next page.

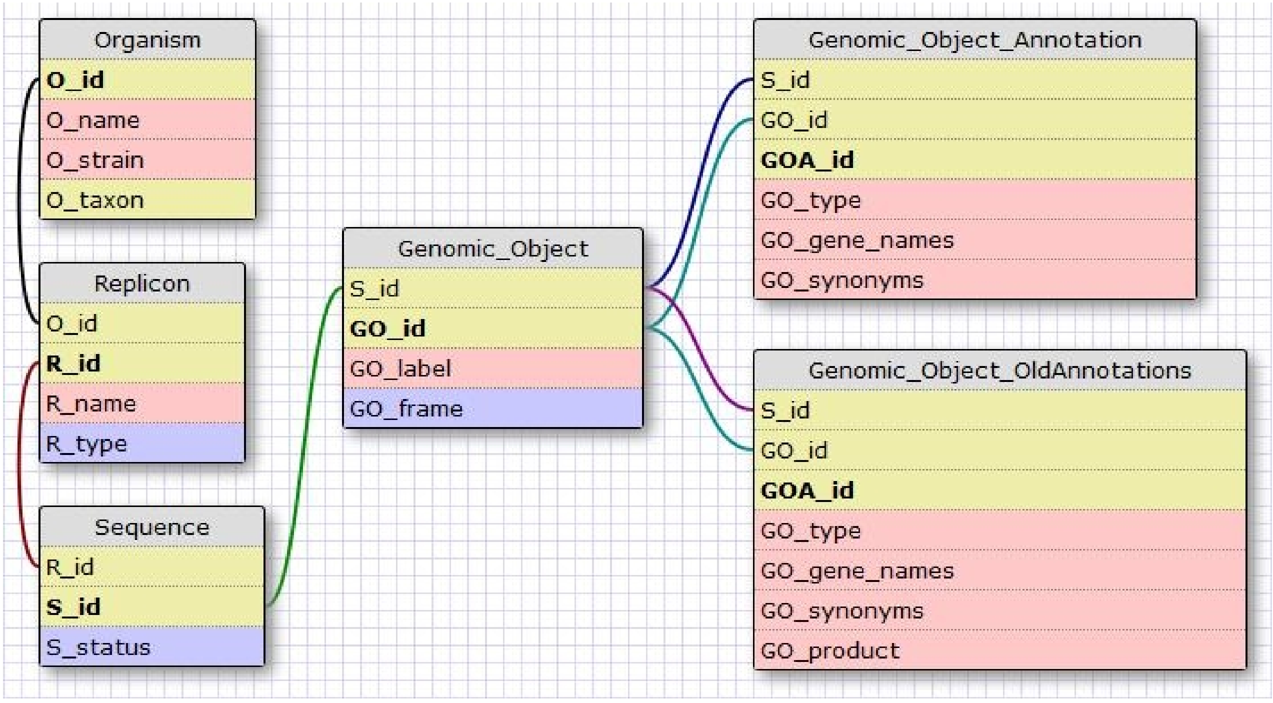


Illustration VII.9: A proposed PkGDB Genomic Object Model

One Genomic Object (identified by a 'GO_id') corresponds to one identified stretch of nucleotides of interest. Annotating this stretch generates a new Genomic Object Annotation (identified by a 'GOA_id'). Any previous Genomic Object Annotation of the same Genomic Object is moved to the Genomic_Object_OldAnnotations table for tracing reasons. As previously, Genomic Objects belong to a given Replicon sequence, which in turn belongs to one Organism. The 'S_id' field is kept in the Genomic Object Annotation tables for quick reference.

On a more technical note, relational databases offer the possibility of using “Foreign Keys”. These describe referential constraints between two tables. They ensure that rows in one table correctly refer to rows in another table. For example, it should not be possible to create a GO for a non-existent Sequence. Another example is of ensuring that if a referred-to row is deleted, then all rows relating to it are also deleted. Basically, foreign keys enforce database coherence and integrity, and also provide a framework on which to build efficient indexes for linking data together. Currently, no foreign keys are used in PkGDB except in some specialised tables. However, it is unclear whether they could be used, as foreign keys are only available for “InnoDB” type tables, and most PkGDB tables are “MyISAM” (a choice motivated by performance issues; the interested reader can find a comparison of the two types at [http://www.kavoir.com/2009/09/mysql-engines-innodb-vs-myisam-a-comparison-of-pros-and-cons.html]). This choice may change if the previously-proposed new model improves performance.

On the same gist, a few tables are not designed correctly in respect to their primary keys. A primary key is a combination of table fields whose values ensure that each row is unique, and is often used is cross-referencing tables. When no such combination exists, it becomes necessary to create an additional column in order to establish a primary key. It can be also interesting to do this if the primary key would otherwise be complex to reference completely from another table (*i.e.* it has multiple columns each with multiple distinct values). Similarly, indexes help speed up table searches for specific field combinations. I have noticed a few tables where the creation of a specific a primary key column was not justified in regard to the other available columns (*e.g.* PkGDB GO_PRIAM_CPD), or where existing indexes were debatable or at least required documentation (*e.g.* MicroCyc Metacyc_Pathway). Reworking these instances would save database space and ensure additional database integrity.

Many of these improvements would be beneficial to the database. However, none have yet been started, as they would require profound adjustments across the MicroScope platform (in PkGDB itself, but also in MaGe and the production system). There are probably other possible improvements that I have not noticed and that require the intervention of a specialist to ensure that MicroScope can keep delivering a high-quality prokaryote genome annotation platform service.

VII.E. Benchmarking

In order to validate the CanOE strategy, I conducted several benchmarking experiments, of which only one is briefly described in the published article and its Supplementary Material. Details for all experiments are given in this section.

VII.E.1. Validating the MinPathLength prior

VII.E.1.a. Protocols

The first part of the CanOE strategy that requires some form of validation our use of the “MinPathlength” (MPL) prior. As described in the paper, we hypothesise that the graph walk-based distances between correctly associated genes and reactions should be shorter than those of incorrectly associated genes and reactions. The gene-reaction or integrated family-reaction scores include an association weighting protocol that uses these distances to favour gene-reaction associations that deal with a gene and a reaction that are “close” in the metabolon.

The idea behind the MPL (as described in the paper's supplementary material) is that it is more likely for genes to be roughly co-linear to their catalysed reactions in a metabolon, rather than the associations be completely random. This idea is illustrated in the figure below. To measure this MPL distance for Known or Potential associations, a path-walking algorithm finds all the shortest paths between the gene to the reaction. Only Known associations can be walked between genes and reactions. If the studied association is Known itself, it cannot be used to walk straight to the reaction, otherwise all Known MPLs would be worth 1 and would not capture local metabolon structure. In order to validate this prior, I conducted an experiment.

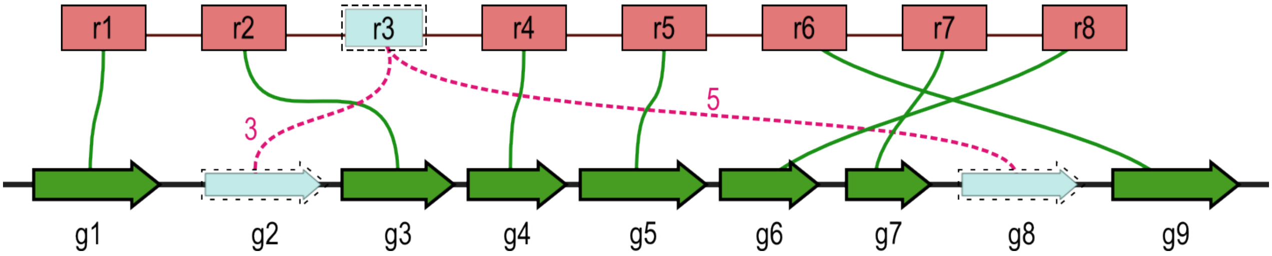


Illustration VII.10: The MinPathLength prior

An example metabolon illustrating our MinPathlength prior. Two gap genes (g2 and g8) are potential candidates for reaction gap r3. Any bioanalyst would consider g2 to be the most likely candidate given the collinear structures in the metabolon. The MinPathLength for each candidate gene is given in pink: g2 is indeed the best candidate according to this measure. Please note that the MPL of gene g8 would not be different if genes g6, g7, g8 and g9 were reversed: the MPL captures local collinearity.

This experiment simply established the distribution of the MPL across all Known gene-reaction associations, in the hope of showing that the distribution is highly skewed towards small values.

VII.E.1.b. Results

The first experiment and its results are included in the Supplementary Material of the article. I show that the distribution does heavily favour short MPLs, though it does not exclude MPLs of up to 10 or more. The distribution itself could have been used to derive a more accurate MPL-weight prior (such as taking quantiles rather than 1/MPL), though I deemed this additional complexity unnecessary in the already complicated CanOE pipeline.

It might be possible to imagine a more rigorous way of verifying this prior. I did not, however, manage to come up with a way of doing this that I thought was implementable in the time left to me for my PhD studies.

VII.E.2. Validating the entire approach

VII.E.2.a. Protocols

The “VerticalOrpheny1” protocol was the first benchmarking protocol I came up with in order to test how well the CanOE method worked and how informative our integrated scores were. Since CanOE deals with identifying candidate genes for orphan reactions, it had to be tested on known cases, *i.e.* on Known gene-reaction associations. The general idea was simple: for a given set of reactions, separately render each one orphan (*i.e.* remove all Known gene-reaction associations concerning that reaction from all organisms), relaunch the CanOE pipeline, and evaluate how well it recovered the removed information in Potential associations and family-reaction scores.

This protocol is calculation heavy, as it involves reproducing the whole CanOE pipeline (primary graph building, metabolon locating, family building, score integration, ranking) for each and every reaction thus “orphaned”. Two choices helped reduce the experimental load. First, we hypothesised that rendering reactions orphans would at worst remove gene families from CanOE data, but not perturb gene family construction; this allowed us to skip the family reconstruction step, using the families built during the normal CanOE run.

We also limited the number of reactions to “render orphan” to those that had at least one in-metabolon Known association in a highly-curated target organism. This typically led us to keeping approximately 340 reactions when using *E. coli* K-12 as a reference. Both these limits (no family recalculation, limited reactions) allowed us to perform the VerticalOrpheny1 benchmarking in a decent amount of time.

Processing the results (metabolons, gene/reaction vertices, gene-reaction associations, gene families, gene/family-reaction association scores) across the several hundred experimental CanOE runs was a complex task as no comparable benchmarking work was available in the literature. I shall present the points deemed important to benchmark below.

Element Recovery: Establishing which metabolons in the experimental runs corresponded to those of the original run was done using Known gene-reaction association sharing (*i.e.* if an experimental metabolon contained at least two same Known gene-reaction associations as a metabolon in the original run, the experimental one was considered a “child” of the original one, thus allowing a one-to-many relationship in the case of metabolon breaking). Using these correspondences, it became possible to track which elements (*i.e.* metabolons, genes, reactions, associations *etc*.) were lost and which were recovered. It was namely possible to estimate the impact of multi-functional genes. Indeed, “orphaning” one reaction from a multifunctional gene leads to the loss of the association (as the gene is still associations with one reaction, it cannot be a candidate for the other), but not of the gene, which can bias gene/family-reaction association scores. Results could be considered across all organisms, or only in the chosen reference organism. Please note that unlike what is described in the following section, association scores and ranks are not considered for this point.

Association Recovery: The objective of the VerticalOrpheny1 benchmarking was to evaluate how well removed Known associations were recovered as Potential associations. Any Potential association can be scored according to gene-reaction association scores or corresponding family-reaction association scores, as described in the paper. Then, for each organism, candidate genes associated to the target “orphaned” reaction can be ranked according to these scores. A cut-off on scores or ranks can be used; any Potential associations kept after cut-off are then positive hits. When Potential associations actually correspond to removed Known associations, these are true associations. It is thus possible to define True Positive, False Positive, and False Negative associations, for varying levels of a cut-off, that can be applied to gene-level or family-level scores or ranks. This is illustrated in the figure below. With these values, it is then possible to draw Precision/Recall curves, which are a handy way of graphically assessing how well a prediction methods performs.

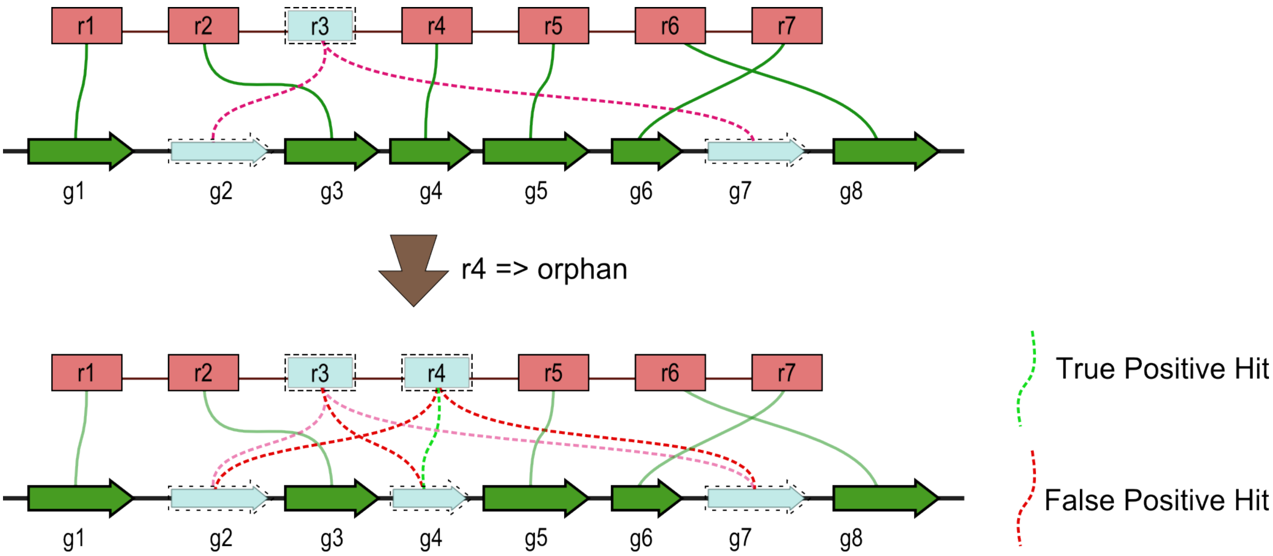


Illustration VII.11: Annotation recovery in CanOE benchmarking

Reaction r4 is rendered orphan by deleting all Known gene-reaction associations involving it. In the presented metabolon, gene g4 catalysed reaction r4. In the new metabolon below, genes g2, g4 and g7 are candidates for r4, and g4 has also become candidate for reaction r3. When not applying any cut-off, amongst the created Potential associations, 1 is a True Positive hit, and 3 are False Positive hits. False Negative hits occur when a gene is not proposed as a candidate for the reaction it catalysed prior to the latter being rendered orphan (due to gene multi-functionality, enzyme subunits, or metabolon partial/total loss).

Findable orphaned reactions: As will be seen, VerticalOrpheny1 benchmarking results were rather catastrophic. Indeed, many metabolons are small, and rendering orphan many reactions leads to the loss of the entire metabolon; reactions close to the extremities of metabolons are also easy to lose. Because of this, it seemed necessary to evaluate benchmarking results on reactions that were deemed “findable”, *i.e.* for which at least one Potential association was recovered (irrespective of it being correct or not). This allows the results to reflect only cases where the CanOE strategy could effectively come up with something, and together with the previous results gives insight into how much information CanOE might actually be missing.

VII.E.2.b. Results

Element recovery: The first figure below gives the distribution of metabolon loss/retrieval cases. Results are pooled across all experiments, for a total of almost 64,000 metabolons. As is rapidly observed, a whopping 50% of those that had originally contained a Known association involving the target reaction lost gene, reaction *and* association after the target reaction was orphaned. 11% lost gene and association (meaning the reaction was still present as a gap), and 21%

lost reaction and association (meaning the gene was still present as a gap). Interestingly, 7% still contained both gene and reaction, but the association was not recovered as a Potential. This corresponds to cases of multi-functional genes (*i.e.* a gene with multiple functions, of which the target reaction, cannot be proposed as a candidate for the target reaction once this one has been orphaned), or multimeric enzymes (*i.e.* a target reaction associated to multiple genes, cannot have candidate genes proposed for it). Finally, only 10% of metabolons actually recovered removed gene-reaction associations as Potentials.

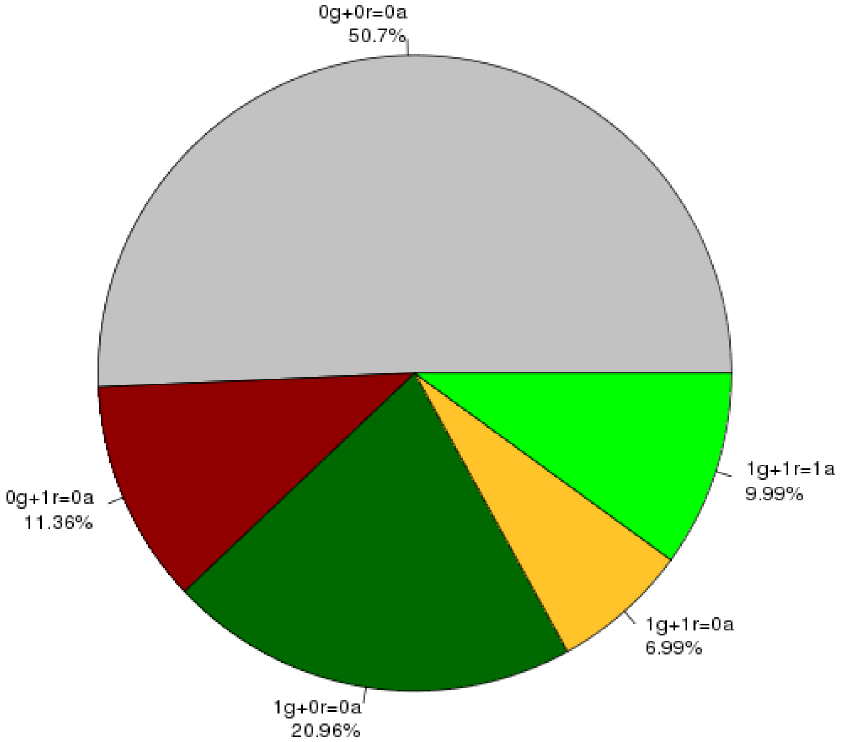


Illustration VII.12: VerticalOrpheny1 element recovery details

Metabolon element recovery/loss details across a total of 63 963 metabolons. During VerticalOrpheny1 benchmarking, the target association can be lost from a metabolon in multiple ways: loss of the gene ("0g" instead of "1g"), loss of the reaction ("0r" instead of "1r"), or simply loss of the association itself ("0a" instead of "1a"). Several cases thus appear: complete loss "0g+0r=0a", gene loss "0g+1r=0a", reaction loss "1g+0r=0a", association loss "1g+1r=0a" (which occurs for enzymes with subunits or multi-functional genes), and finally association recovery "1g+1r=1a".

When only considering findable reactions (see below), things improve somewhat. Most importantly, metabolons containing retrieved gene-reaction associations now represent 30% of the total, which is now only about 16,000 metabolons. As expected, complete losses have decreased, now affecting 36% of metabolons (not 0%, as the definition of "findable" does not remove all metabolons with lost associations, only those whose target reaction had no recovered associations anywhere).

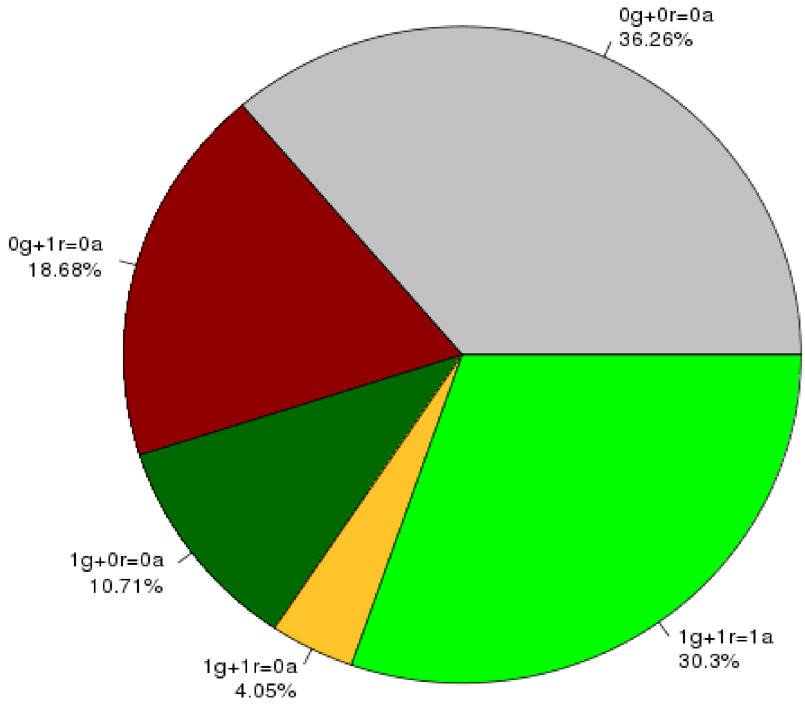


Illustration VII.13: VerticalOrpheny1 findable element recovery details

Metabolon element recovery/loss details across a total of 16,082 metabolons. During VerticalOrpheny1 benchmarking, the target association can be lost from a metabolon in multiple ways: loss of the gene ("0g" instead of "1g"), loss of the reaction ("0r" instead of "1r"), or simply loss of the association itself ("0a" instead of "1a"). Several cases thus appear: complete loss "0g+0r=0a", gene loss "0g+1r=0a", reaction loss "1g+0r=0a", association loss "1g+1r=0a" (which occurs for enzymes with subunits or multi-functional genes), and finally association recovery "1g+1r=1a".

These results encourage us to use the "findable" filter for further results.

Gene-level recalls and ranks: As the figures below show, results are rather catastrophic. Indeed, the highest recall value is not even 20%; the best precision is 55% (for a recall of roughly 15%) obtained for the G2R_W score, at a minimum cut-off of 0.7. Clearly, gene-level scores and ranks are of little interest when considering the CanOE strategy as a whole.

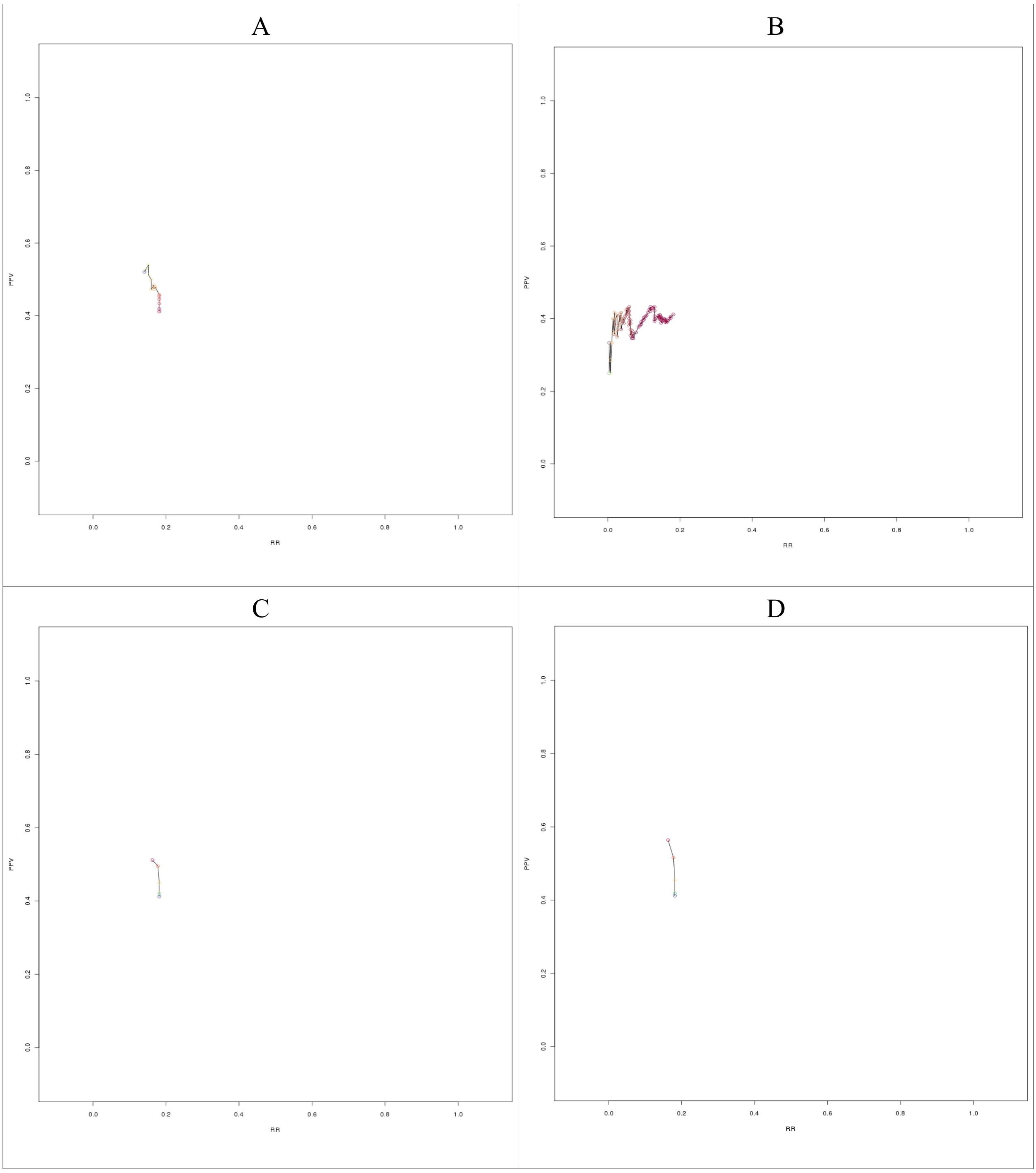


Illustration VII.14: VerticalOrpheny1 gene-level benchmarking results

Precision-Recall curves for a varying cut-off applied to numeric scores (A, B) or to the ranks thereof (C, D), for gene-to-reaction (A, C) and reaction-to-gene (B, D) scores. Most lenient cut-offs correspond to the highest Recall values.

Family-level recalls and ranks: The curves below are slightly better than the corresponding gene-level ones. Especially R2F_W score (precision of 62%, recall of 15%, for a minimum score cut-off of roughly 0.3) and R2F_W rank (precision of 61%, recall of 15%, for a maximum rank cut-off of 1). Still, values remain low and precision/recall trade-off for varying values of cut-off is uninteresting.

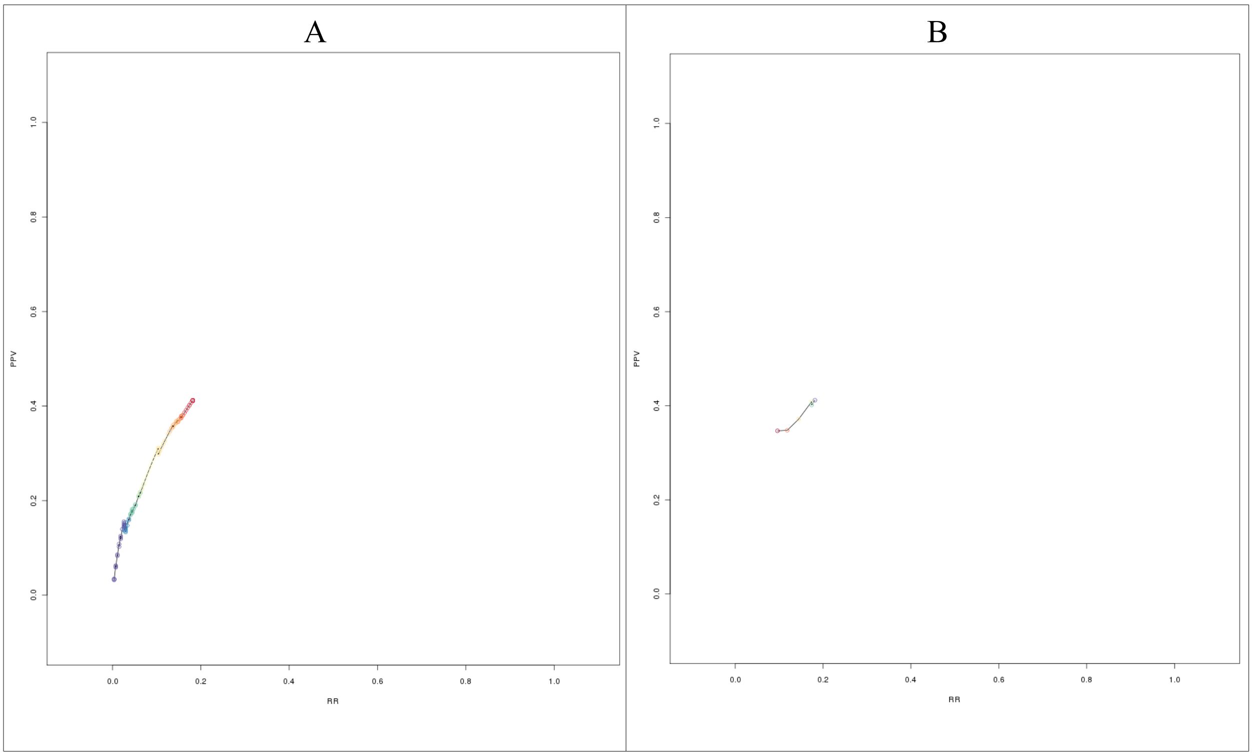


Illustration VII.15: VerticalOrpheny1 family-level benchmarking results for Coverage score

Precision-Recall curves for a varying cut-off applied to numeric Coverage score (A) or to the rank thereof (B).

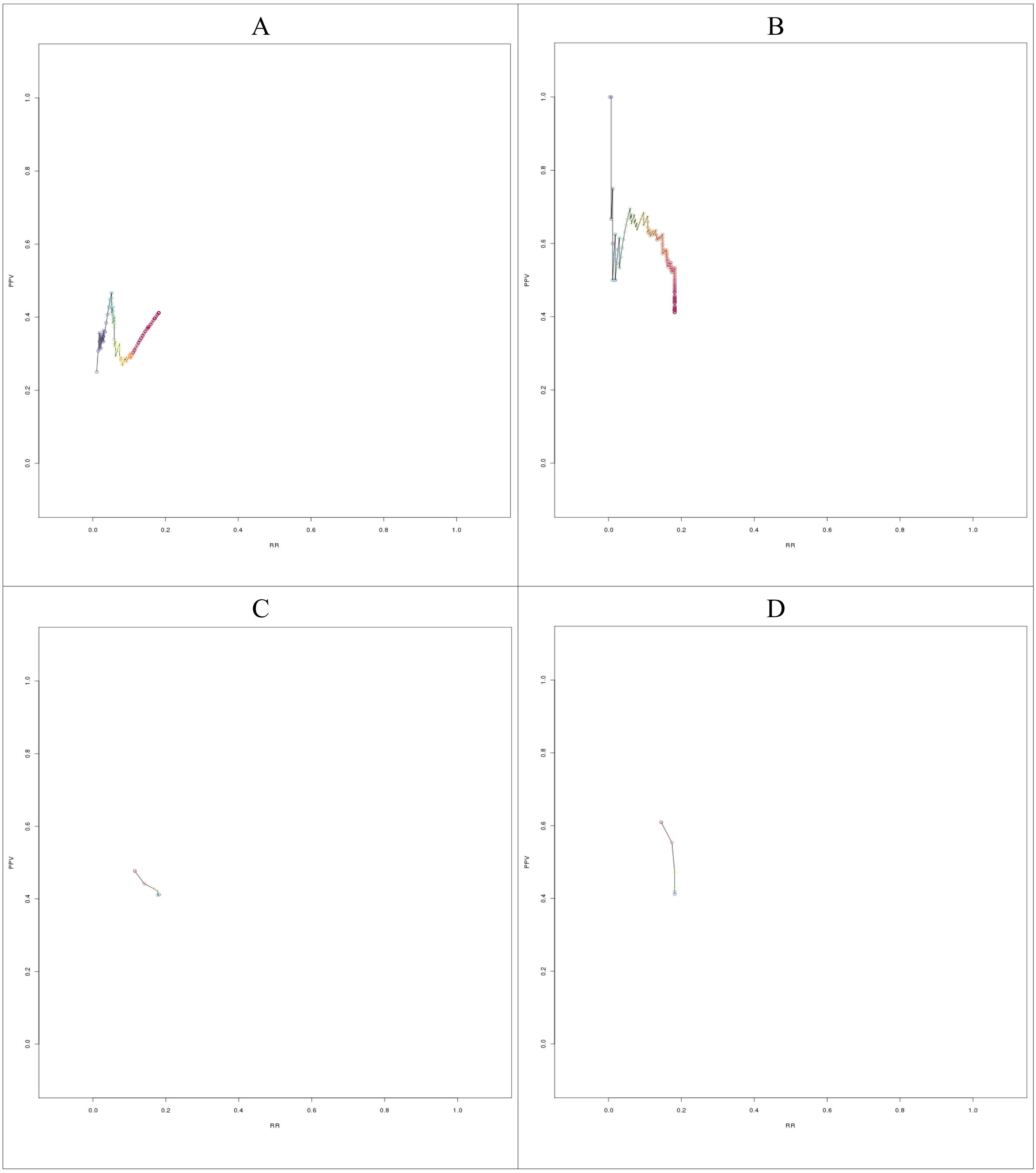


Illustration VII.16: VerticalOrpheny1 family-level benchmarking results

Precision-Recall curves for a varying cut-off applied to numeric scores (A, B) or to the ranks thereof (C, D), for family-to-reaction (A, C) and reaction-to-family (B, D) scores.

We shall thus now examine the corresponding benchmarking results for “findable” reactions only. These are given in the following figures.

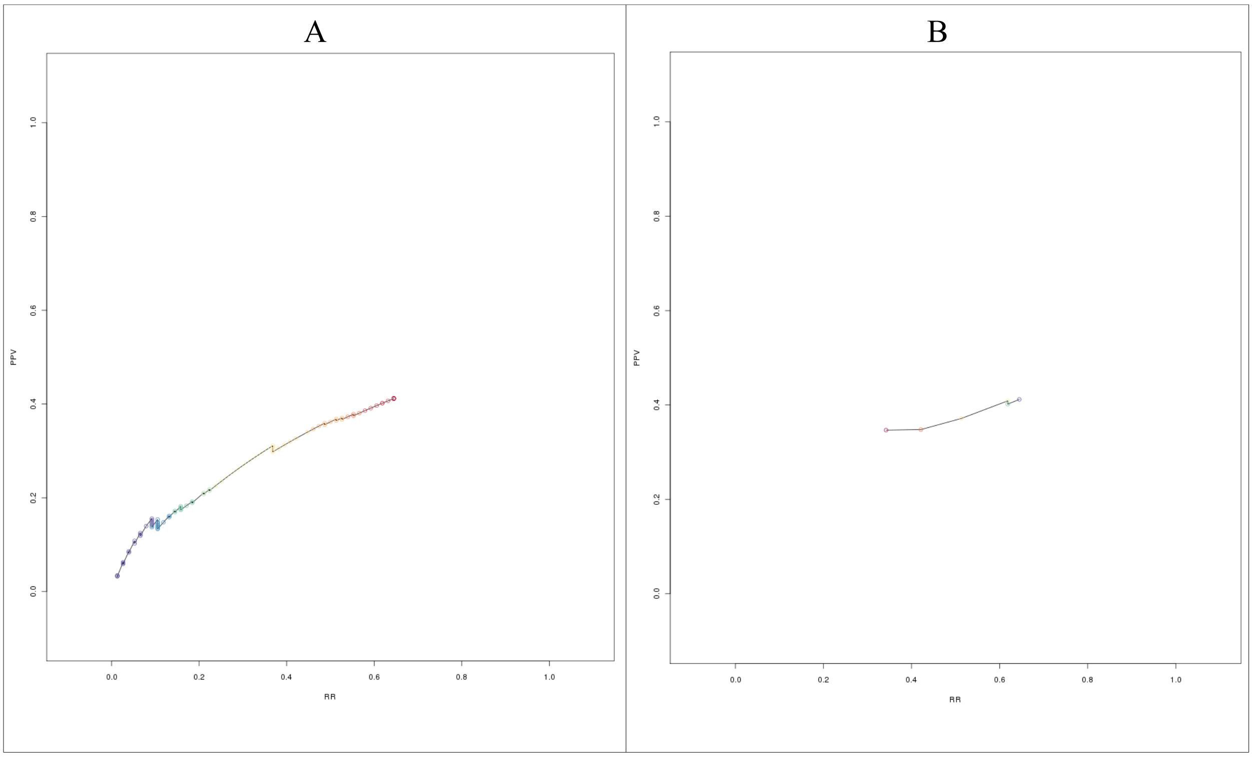


Illustration VII.17: Findable VerticalOrpheny1 family-level benchmarking Coverage results

Precision-Recall curves for a varying cut-off applied to numeric Coverage score (A) or to the rank thereof (B), when only considering findable reactions.

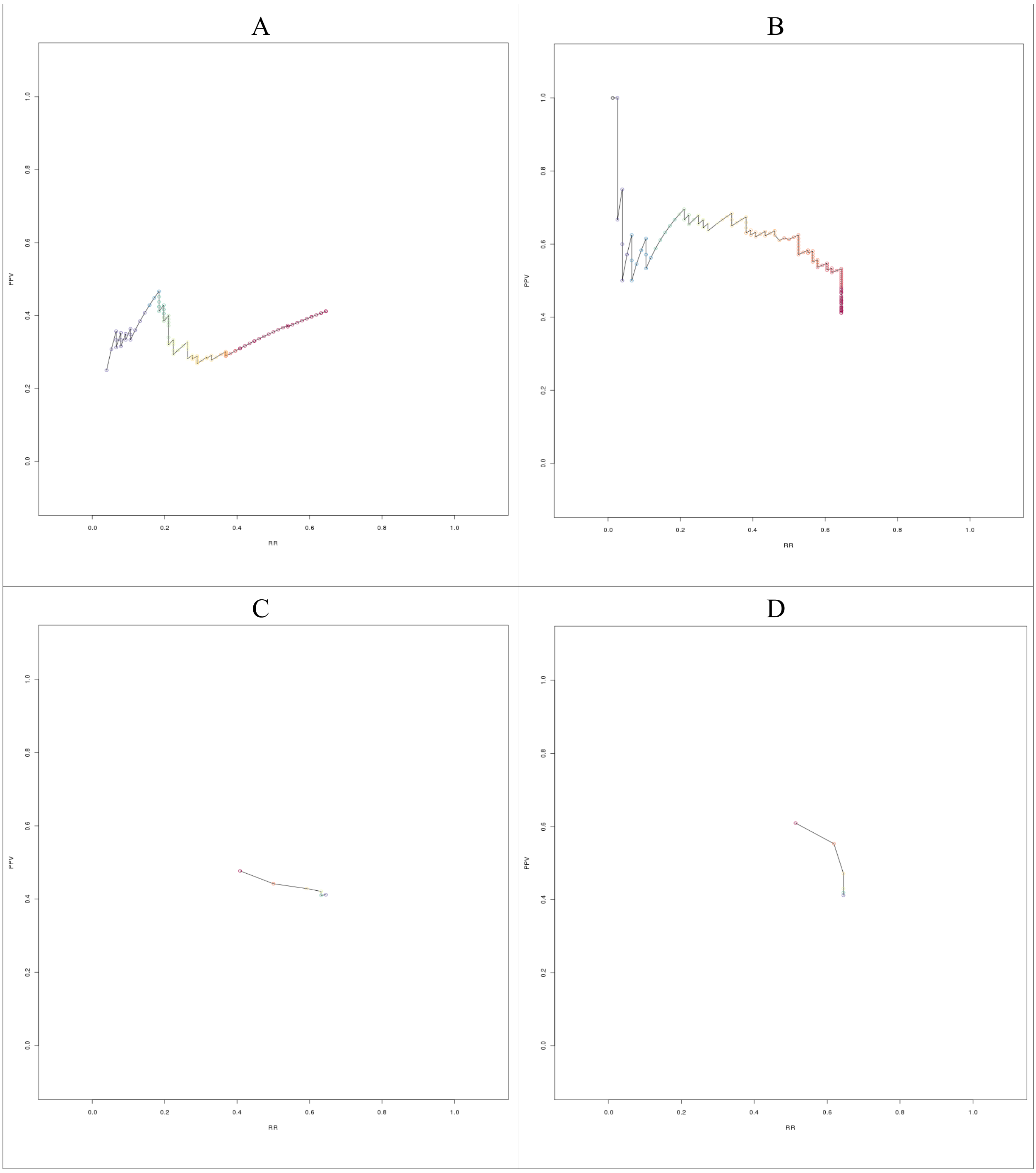


Illustration VII.18: VerticalOrpheny1 findable family-level benchmarking results

Precision-Recall curves for a varying cut-off applied to numeric scores (A, B) or to the ranks thereof (C, D), for family-to-reaction (A, C) and reaction-to-family (B, D) scores, when only considering findable reactions.

These results are much more encouraging than the previous, illustrating that the CanOE strategy is relatively powerful at recovering gene-reaction associations, at least when metabolon structures allow it. Indeed, for the ranks based on the F->R score, we obtain roughly 40% precision at 63% recall when we keep all results, which improves to 57% precision at 61% recall when keeping only the 1st and 2nd best ranking candidate genes. Transposing to real-life cases, we can expect CanOE to be powerful for associations involving prokaryote orphan reactions that have at least several associations with a given gene family, though it very likely misses large portions of metabolism that just cannot be seen with current levels of genome annotation.

In conclusion, benchmarking results with the VerticalOrpheny1 protocol were relatively poor, and bad results for the most part were due to metabolon loss because of known association removal ("metabolon breaking"). This encouraged us to develop another protocol, using a more favourable approach, that could still allow us to serenely conclude how informative our integrative scores were.

VII.E.3. Validating the integration over n only

VII.E.3.a. Protocol

The "VerticalOrpheny2" protocol is the one described in the article. It is similar to VerticalOrpheny1, except that a) **it does not recalculate metabolons** and b) does not use reactions from a reference organism, rather a selection of reactions respecting the following rule: only select reactions that have at least one known association in a metabolon that is "big enough" in one organism. A "big enough" metabolon is one that would not automatically be lost after removal of the known association (*i.e.* contains at least 3 reactions, 3 genes and 3 known associations). The rationale behind this is that metabolons are the given information and their localisation is deterministic; what it is important to test is how the whole strategy combines them into informative results. Skipping metabolon recalculation denies metabolon loss[16], thus ensuring that CanOE works with all known information, less the actual orphaned reaction associations. Though heavily biased, this simpler approach allows us to evaluate the integration over multiple genomes procedure.

VII.E.3.b. Results

The results obtained with this benchmarking approach are already discussed in the paper and are rather reassuring. Indeed, maximum recall is slightly more than 80% for all scores, for a precision of roughly 52%. Filtering out worse-ranked candidates can improve precision at the cost of recall; the best trade-off is obtained for the ranks based on the R=>F score, as shown in the article. Finally,

[16] It is similar in this way to the "findable" condition in the VerticalOrpheny1 benchmarking.

we can observe that in all cases, precision and/or recall are higher for a given rank threshold on the integrated scores (F=>R and R=>F) rather than the gene-level scores (G=>R and R=>G). This confirms the intuition that integrating over many organisms can improve results.

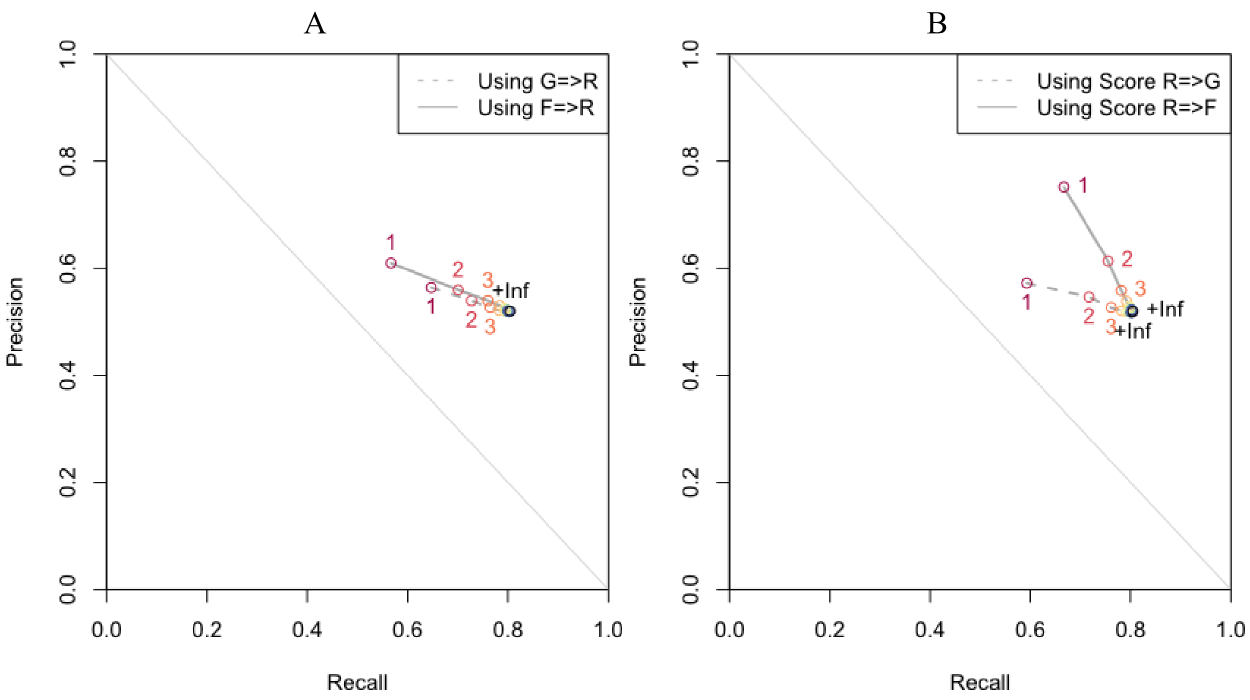


Illustration VII.19: VerticalOrpheny2 benchmarking results

A: Precision-Recall curves using gene-to-reaction (G=>R) or family-to-reaction (F=>R) ranks. B: Precision-Recall curves using reaction-to-gene (R=>G) or reaction-to-family (R=>F) ranks.

VII.F. Web Interface overview

All the results of the CanOE strategy need to be made accessible to the bioanalyst users of the MicroScope platform if they are to be of any use. CanOE has been integrated into MicroScope's web interface, MaGe, to this end. I shall thus present MaGe, before detailing what the CanOE interface has become over the past months.

VII.F.1. The MaGe web interface

MaGe (Magnifying Genomes) is the web-based interface to the tools and data of the MicroScope platform [4]. It is mostly coded in PHP and uses AJAX-based technologies to create dynamically generated and interactive web pages, with asynchronous creation and submission of SQL queries for retrieving required data (*i.e.* that can be sent on demand even after a page is rendered by the web browser). Page structure itself is coded using CSS and HTML. Recent work by Gregory SALVIGNOL has lead to the development of MaGe version 2.0, which uses increasing amounts of dynamic technologies, and aims to meet common web page quality and coding standards (HTML 5.0, W3C standards...).

The interface is organised into 3 types of pages. The first of these is the Genome Browser, that represents the location of Genomic Objects across all 6 translation frames of a given DNA sequence chosen by the user. A homology/synteny viewer below allows rapid appreciation of the conservation of one or multiple gene sequences between the sequence of interest and other selected genomes (from related or distant organisms). Finally, a dynamic table lists interesting data for all currently visible GOs. With this page, a user bioanalyst can rapidly navigate the genome, appreciating conservation with others, and accessing basic annotation.

The second type of page is the Genomic Object Editor, which is accessible (amongst others) from the Genome Viewer. It allows consultation of all annotation data (automatic predictions, various bioinformatic method results, previous manual annotations, GO sequence, annotation history...) for a given Genomic Object. This is also the interface that bioanalysts can use to create new annotations when they have sufficient editing rights.

The final type of page is the "template" tool page. A large number of bioinformatic tools are available, each coming with their own interface tailored to their needs. Proposed services include BLASTing, genome-wise descriptive statistics, specific annotation queries, specialised interfaces for genome projects with particular requirements, gene cart handling and queries, various comparative genomics and metabolic tools, and more. The diversity of approaches offered to the bioanalyst is one of MicroScope's strengths.

Additional administrative, user preferences, news pages, etc., that do not concern direct interfacing with MicroScope data, are also available.

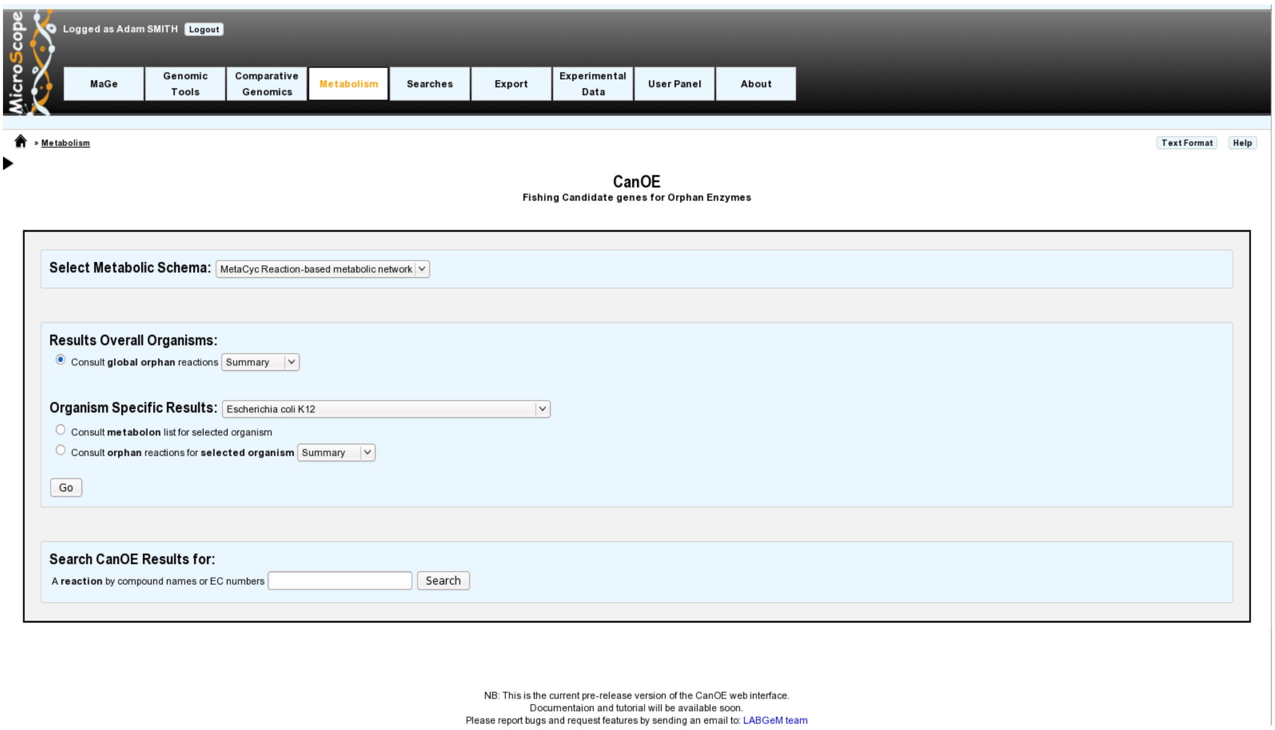
VII.F.2. The CanOE web interface

Only a pre-release version of the web interface designed to make CanOE results available to the scientific community was ready for the publication of the CanOE article; it was thus decided that the final version would be prepared and published in the following MicroScope platform-wide publication. I designed this version on the basis of a previous draft version that I had developed in order to ease my exploration of CanOE results. David VALLENET and Gregory SALVIGNOL integrated it into MaGe. I shall present this work and its perspectives, as they are relevant to my thesis and the professional career I am planning.

The MicroScope platform is designed to help expert bioanalysts functionally annotate genes in prokaryotic genomes. The CanOE results thus have to be made accessible to them in the clearest manner, presenting them with precise biological questions in mind. Several “access points” have been imagined for these results, some of which have already been implemented in the beta version. Current implementation accounts for the choice of metabolic schema, though only the MetaCyc-based schema is currently used.

VII.F.2.a. CanOE main page

CanOE can be used as a stand-alone tool for the MicroScope platform at [www.genoscope.cns.fr/agc/microscope/metabolism/canoe.php]. It can thus be accessed, like any other tool, via a MaGe menu (though, until the beta version is properly put into production, the menu item will remain hidden from the platform users). The main page is basically an interface for requesting CanOE data. It allows users to select the metabolic schema, to consult a list of metabolons for a specific organism, to consult lists of local or global orphan reactions with candidates at different detail levels, or to search CanOE for genes or reactions. Available results are obviously limited to organisms for which the user has access rights. The results that can be obtained from this interface are described in the following sections.



CanOE
Fishing Candidate genes for Orphan Enzymes

Select Metabolic Schema: MetaCyc Reaction-based metabolic network

Results Overall Organisms:
● Consult **global orphan** reactions Summary

Organism Specific Results: Escherichia coli K12
○ Consult **metabolon** list for selected organism
○ Consult **orphan** reactions for **selected organism** Summary
Go

Search CanOE Results for:
A **reaction** by compound names or EC numbers Search

NB: This is the current pre-release version of the CanOE web interface.
Documentaion and tutorial will be available soon.
Please report bugs and request features by sending an email to: LABGeM team

Illustration VII.20: CanOE main web page

CanOE's main page can be accessed from the "Metabolism" MicroScope tools menu. Most of the access points imagined for the CanOE tool are available from this page: listing metabolons per organism, consulting local or global orphan reactions at varying levels of detail, and searching trough the results with keywords.

VII.F.2.b. Organism-centric view

Most bioanalysts work on the annotation of a single genome (or a group of related genomes). It is thus be interesting for them to be able to quickly access the metabolons of their target organism(s).

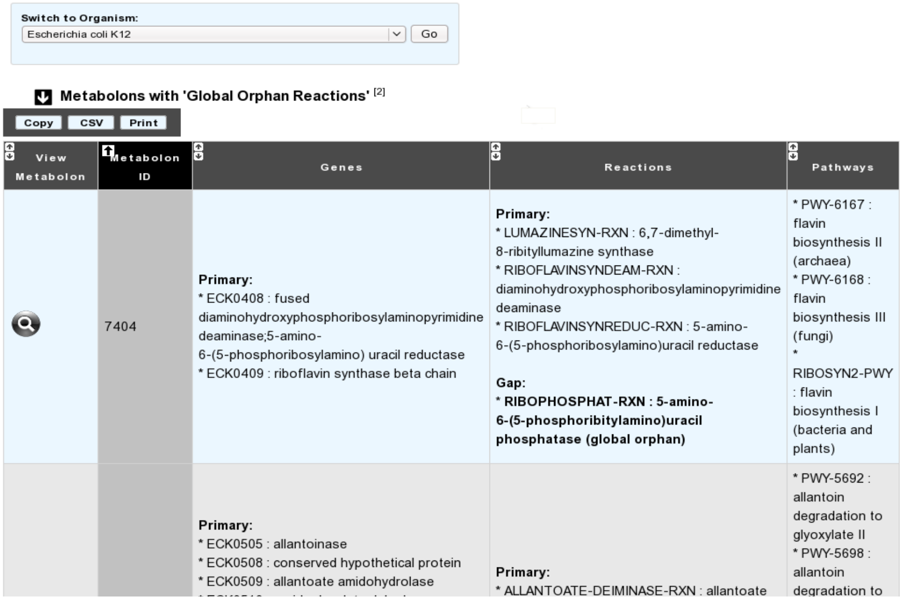
With this focus in mind, three access points have been imagined:

- **Genome viewer (imagined):** create a graphical representation of metabolons along the genome, inspired by the current presentation of syntenies along the genome. Genes belonging to a same metabolon would have illustrative copies of a same colour in a "metabolon track", which could be clicked to open the metabolon viewer.
- **Gene editor (imagined):** the Gene Editor lists all the MicroScope data available for a given gene in various tables, including synteny belonging. It would be trivial to add a Metabolon table to the list of available data for a given gene, especially as genes rarely belong to more

than one metabolon[17] (94.5% single metabolon, 4.7% two metabolons, 0.9% above).

- **Metabolon List (implemented):** Simply list all the metabolons detected within a target organism, allowing a user to scan through the structured metabolic knowledge located by the CanOE strategy.

The Metabolon List results page (accessible from the CanOE main page) separates metabolons into categories based on their reaction content: 1) metabolons with at least one global orphan reaction, 2) metabolons with at least one local orphan reaction and no global orphan reactions, 3) metabolons with at least one gap reaction but no orphan reactions, and 4) "complete" metabolons in respect to reactions. Each metabolon is described in terms of primary/gap genes and reactions, as well as associated metabolic pathways, and gives access to the MetabolonViewer page in order to graphically represent it. The other organism-centric access points will be implemented in the future.



Metabolons with 'Global Orphan Reactions' [2]

View Metabolon	Metabolon ID	Genes	Reactions	Pathways
🔍	7404	**Primary:** * ECK0408 : fused diaminohydroxyphosphoribosylaminopyrimidine deaminase;5-amino-6-(5-phosphoribosylamino) uracil reductase * ECK0409 : riboflavin synthase beta chain	**Primary:** * LUMAZINESYN-RXN : 6,7-dimethyl-8-ribityllumazine synthase * RIBOFLAVINSYNDEAM-RXN : diaminohydroxyphosphoribosylaminopyrimidine deaminase * RIBOFLAVINSYNREDUC-RXN : 5-amino-6-(5-phosphoribosylamino)uracil reductase **Gap:** *** RIBOPHOSPHAT-RXN : 5-amino-6-(5-phosphoribitylamino)uracil phosphatase (global orphan)**	* PWY-6167 : flavin biosynthesis II (archaea) * PWY-6168 : flavin biosynthesis III (fungi) * RIBOSYN2-PWY : flavin biosynthesis I (bacteria and plants)
		Primary: * ECK0505 : allantoinase * ECK0508 : conserved hypothetical protein * ECK0509 : allantoate amidohydrolase	**Primary:** * ALLANTOATE-DEIMINASE-RXN : allantoate	* PWY-5692 : allantoin degradation to glyoxylate II * PWY-5698 : allantoin degradation to

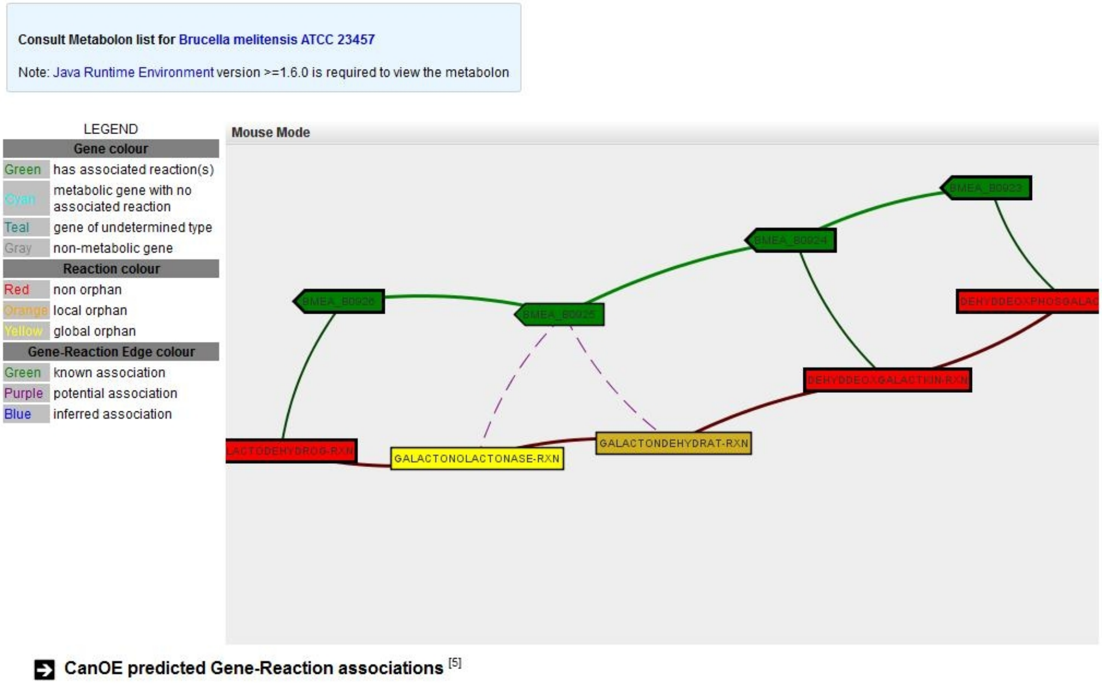
Illustration VII.21: CanOE metabolon list

The MicroScope user can access a list of metabolons for a specific organism to which he/she has access. Metabolons are described in terms of gene and reaction content, associated metabolic pathways, but foremost, by any local/global orphan reactions they may contain.

[17] Indeed, a single Known gene-reaction association can only belong to a single metabolon. Genes belonging to multiple metabolons thus have multiple Known gene-reaction associations, each captured by a different metabolon.

VII.F.2.c. Metabolon Viewer

The first and foremost of representations of metabolon data is obviously a graphical illustration of the metabolon itself. To this end, I implemented a Java package (based on the Jung package [http://jung.sourceforge.net/]) that could draw a metabolon using CanOE results extracted from the MicroScope database. Current implementation displays genes as arrows and reactions as rectangles, with colour variations corresponding to specific flags (annotated/non-annotated gene, metabolic/non-metabolic gene, global/local/non orphan reaction), and edges with confidence-specific patterns (Known/Potential/Inferred). I intend to represent compounds in a further version, in order to increase readability of the metabolic part. Ideally, the genes and reactions should be laid out automatically to best highlight the metabolon structure. I was not, however, able to discover an algorithm for efficiently doing this low-priority task, and nodes currently must be manually displaced by the user. Alain VIARI has helped me find a promising algorithm that is implemented in Javascript rather than Java. I shall use it to replace the currently too-heavy MetabolonViewer applet when I have the time.



Consult Metabolon list for Brucella melitensis ATCC 23457

Note: Java Runtime Environment version >=1.6.0 is required to view the metabolon

LEGEND

Gene colour	
Green	has associated reaction(s)
Cyan	metabolic gene with no associated reaction
Teal	gene of undetermined type
Gray	non-metabolic gene
Reaction colour	
Red	non orphan
Orange	local orphan
Yellow	global orphan
Gene-Reaction Edge colour	
Green	known association
Purple	potential association
Blue	inferred association

CanOE predicted Gene-Reaction associations [5]

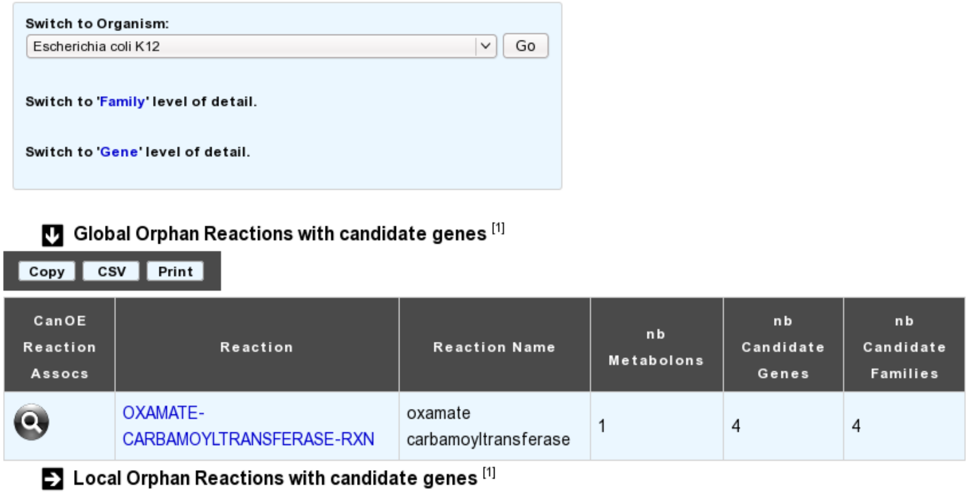
Illustration VII.22: CanOE metabolon viewer

The MetabolonViewer generates a graphical representation of a metabolon, with genes, reactions, and all types of edges. Users can currently move each vertex manually in order to improve the graph's layout. Basic information is available by right-clicking on the vertices; full information about metabolon contents is given in the tables below the viewer itself.

The MetabolonViewer page contains additional information, such as the colour legend, the list of genes, reactions and metabolic pathways with details, the list of metabolon gene-reaction associations (Known, Potential and Inferred), as well as a list of any other metabolons which share genes with the current one (useful for overlapping metabolons or alternate pathway metabolons).

VII.F.2.d. Orphan reaction pages

Some bioanalysts might approach MicroScope with the hope of finding candidate genes for local or global orphan reactions. In order to assist them in this endeavour, we propose a web page listing all local or global orphan reactions with at least one candidate gene in at least one metabolon. The results can be consulted at three different detail levels: a summary of all orphan reactions with candidates, a family-level detail of candidate families for each reaction, or a gene-level detail of candidate genes for each reaction. According to the chosen level, results provide quick link access to data pages describing the reactions, the families, the genes, the metabolons, and the family-reaction associations involved.



Switch to Organism: Escherichia coli K12

Switch to 'Family' level of detail.

Switch to 'Gene' level of detail.

Global Orphan Reactions with candidate genes [1]

CanOE Reaction Assocs	Reaction	Reaction Name	nb Metabolons	nb Candidate Genes	nb Candidate Families
🔍	OXAMATE-CARBAMOYLTRANSFERASE-RXN	oxamate carbamoyltransferase	1	4	4

Local Orphan Reactions with candidate genes [1]

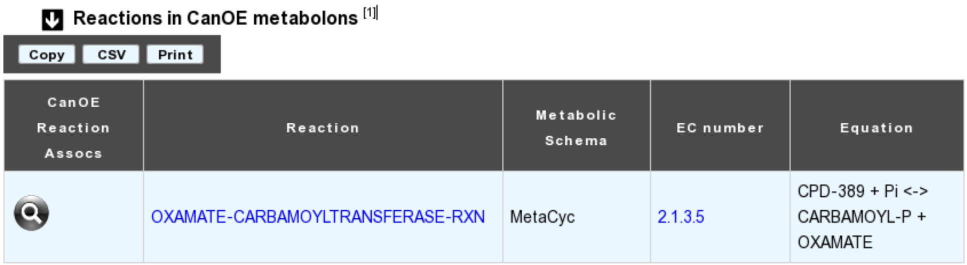
Illustration VII.23: CanOE orphan reaction pages

The orphan reaction pages can specify, at three levels of detail (summary, family, and gene), CanOE's candidate genes for either local or global orphan reactions, with dynamic links to other relevant pages.

VII.F.2.e. Search result page and object-specific pages

Searching for a gene returns results and links concerning genes, as well as families and metabolons containing those genes, for all metabolic schemas. Searching for a reaction returns the list of all CanOE reactions that contain the specified keyword in their name or equation, for all metabolic schemas.

Each object used in the CanOE data model (reaction, family, metabolon, family-reaction association) has its own description page (the metabolon page being the already-described MetabolonViewer page). These provide internal links to other associated objects and results, as well as external links for metabolic reactions and pathways (current implementation works for MetaCyc).



Reactions in CanOE metabolons [1]

CanOE Reaction Assocs	Reaction	Metabolic Schema	EC number	Equation
🔍	OXAMATE-CARBAMOYLTRANSFERASE-RXN	MetaCyc	2.1.3.5	CPD-389 + Pi <-> CARBAMOYL-P + OXAMATE

Illustration VII.24: CanOE search result page

A user can search for reactions or genes using keywords. A reaction search locates reactions in all metabolic schemas with the specified keyword associated to them (EC numbers, compounds, part of reaction name...). A gene search locates genes and their containing families with GO_labels or gene names including the keyword. For each type of search, a table is returned with a list of hyperlinks to relevant CanOE objects. For security reasons, the gene search is not intended to be made publicly available.

Altogether, these web pages should provide sufficient access to CanOE results for bioanalysts to exploit them correctly in the manners discussed below. Documents describing the tool and tutorials for bioanalyst users of MicroScope are under preparation.

VII.G. CanOE uses

The CanOE strategy, once its development is complete, shall offer bioanalyst users of the MicroScope platform 4 services, graded in increasing usefulness and decreasing result volumes:

1. CanOE predicts metabolons, which are original objects that can be used to **visualise** and synthesise metabolic and genomic information, easing bioanalysis;
2. These metabolons serve as a basis for **proposing Potential gene-reaction associations**, with **scores** reflecting results integrated across several genomes and thus conservation;
3. Combining Known and Potential gene-reaction associations across sequence similarity-based families allows the **creation of relatively high confidence Inferred associations** that bioanalysts may use to complete gene annotations;
4. Finally, metabolons can propose **candidate genes even for sequence-orphan enzymatic activities**, as association generation is not sequence-dependent.

All 60,000[18] metabolons are available as support for point 1. Only a sub-selection of them (3,867) contain novel association propositions necessary to point 2. Only 1,125 of these metabolons contain Inferred associations (point 3). Finally, despite the effort put into the design of CanOE, a disappointing number of metabolons (597) contain one of the 78 sequence-orphan activities that have candidate genes (point 4). This last point is discussed later, in section VII.H.2.

Thankfully, the strategy is open-ended, in that many improvements can be made to it, that hopefully will increase the number of interesting predictions. Furthermore, several other use cases for metabolons can be imagined. Finally, it may one day be able to take enzymatic promiscuity into account. I shall discuss these points below, before concluding the main project of my thesis.

VII.H. Discussion and perspectives

VII.H.1. CanOE general improvement leads

The VerticalOrpheny1 benchmarking protocol showed that CanOE can be powerful for orphan reactions that have been seen in a few metabolons. In this sense, we feel that it is possible to defend the “candidate gene finding for orphan enzymes” aspect of CanOE. Obviously, the small number of orphan reactions for which metabolons or candidate genes are available is disappointing. Furthermore, bioanalysis of the propositions leads one to consider many as false positives. We argue

[18] Numbers are given for August, 2011

that if bioanalysis of these propositions allows the identification of only a few correct candidate genes for orphan enzymes, then that is better than nothing; furthermore, as pointed out in the previous conclusions, CanOE results are interesting to the bioanalyst for other, non orphan-dependant reasons. We also argue that improvements to a) CanOE metabolic networks, b) CanOE gene family construction (or use of gene similarity), c) CanOE functional dependence indicators, and d) MicroScope annotations will alleviate this. Indeed, point a) will increase the number of available reactions and better connect them to each other (even possibly involving transport activities); point b) might help refine association scores; point c) will make large, more process-related metabolons possible; point d) will allow CanOE to capture additional metabolic contexts, thus increasing possibilities for hypothesis generating. All these improvements will be discussed amongst the following sections.

VII.H.2. Improving the metabolic graph

As pointed out in the article, a not-so-small fraction of orphan enzymatic activities are excluded from the global metabolic graph used in CanOE's Metaboloniser because they cannot be connected to other metabolic reactions by only keeping “main” compounds. In the MetaCyc schema presented in this thesis and in the article, this could be due to the absence of metabolic pathways including these orphan reactions, or because the containing pathways are too small or not linked to others in a way that allows the orphan reaction to be linked to other reactions in other pathways. In the KEGG-based metabolic schemas presented in this manuscript, reaction disconnection could just as well be due to the absence of a parent metabolic pathway, to overly-stringent or sub-optimal filtering conditions. In any case, disconnected orphan reactions cannot be included in a metabolic context, and *de facto*, cannot be found in a metabolon. Improving the metabolic graph in any way would seem the best place to start in order to capture additional orphan reactions, and hopefully, more candidate genes.

The source metabolic databases are in a state of constant manual curation, MetaCyc with in-house curators and its user feedback, KEGG with the research projects piloted by Kanehisa labs, both being fuelled by the novel experimental discoveries described in published literature. The metabolic network can thus only improve in terms of coverage, though should connectivity be an issue (as was the case with the KEGG schemas), then improved protocols for defining “main” compounds might be necessary, using for example better KEGG Map data, RPAIR data, or even simple Reaction data.

Another possibility, discussed more in length in chapter X, would be the use of CanOE and

BKACE-like (see chapter VIII) strategies in an iterative way, gradually exploring further the metabolic space of prokaryote organisms.

Finally, a bibliographical effort may be useful. I detail this approach in the following section.

VII.H.3. New Enzymatic Activity Survey

During this thesis and more specifically during the genesis of CanOE, we pondered on how the LABGeM could contribute to reducing the number of orphan enzymes in a way that would readily benefit the MicroScope platform. As has already been suggested (see section VI.B and [2]), a non-negligble fraction of orphan enzymatic activities may be orphans only because of insufficient efforts in the computerisation of knowledge already present in scientific articles. Also, the number of known activities continues to grow each year (several hundred EC numbers gained over the past six months), with the creation of candidate activities (validated or not), which further complicates the tracking of reaction bibliography. The recent UniProt effort [3] helped reduce by a few percent the number of global orphan enzymes, but many still seek parent genes.

In order to deal with these shortcomings, we imagined setting up an internal bibliographical tool called “New Enzymatic Activity Survey” (NEAS) at the Genoscope that would allow us to follow the evolution of enzymatic activities (*i.e.* discovery, formalisation, modification, transfer/renaming, deletion,...), their annotation status (*i.e.* how many genes, across how many organisms, are annotated with such and such activity?), and would also allow us to create new, user-defined activities and protein-activity annotations backed by bibliographic evidence. The use of such a tool would lead to :

- a reduction in artefactual orphan activities
- the maintenance of an up-to-date, manually curated metabolic resource
- the maintenance of an up-to-date, manually curated bibliographical resource for:
    - the discovery of novel activities
    - their assignment to gene/protein sequences

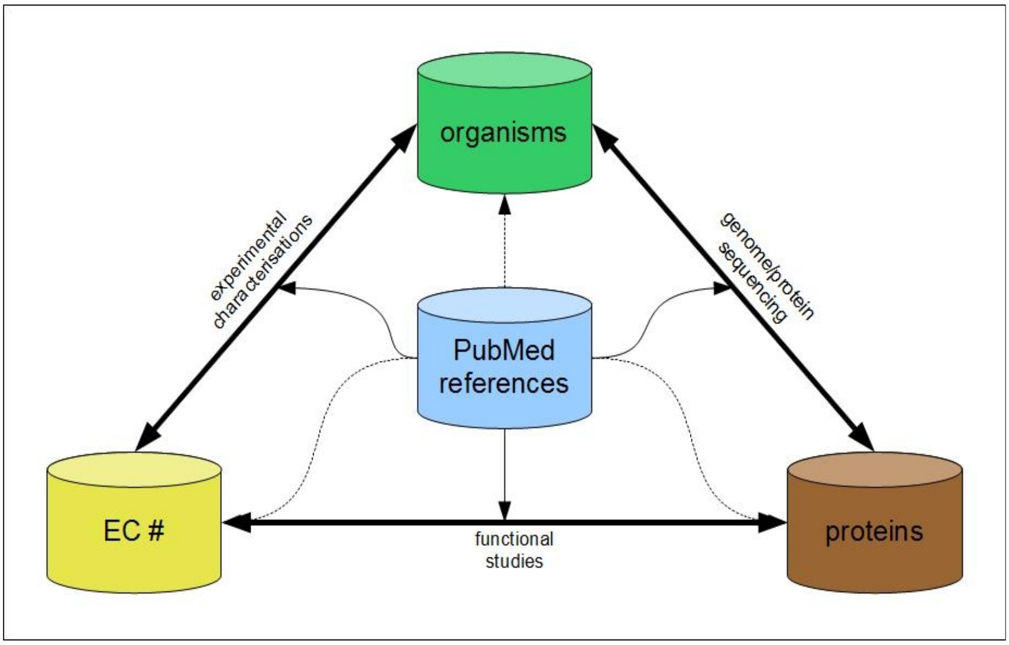


Illustration VII.25: Schema of NEAS objects and their relationships

NEAS would be dedicated to tracking the associations between metabolic reactions (formalised by EC number for example), genes/proteins, and their sequences and host organisms. All of these objects and associations would require references to scientific literature. By keeping such a resource up to date, NEAS would provide precise tracking of activities.

All in all, automated and manual gene functional annotation methods would all benefit from this kind of tool. It would also contribute to establishing descriptive statistics of the field over time, such as described in section VI.B. However, due to time and resource constraints, this project was not initiated during my thesis, though it should see the light in the coming months. When it is ready, NEAS should be able to improve CanOE results by feeding it better coverage of prokaryote metabolism.

I shall now discuss the other improvements or ideas we have imagined that could be applied to the CanOE strategy.

VII.H.4. Various CanOE improvements

VII.H.4.a. Annotation transfer by Context Similarity

Our first musing concerned **how could similar metabolons be used to transfer annotations between non-similar genes**. This idea is illustrated in the figure below. The reasoning is thus: if a metabolon M1 shares many reactions with metabolon M2, and several of these reactions are catalysed by genes that have detectable sequence similarity between M1 and M2, should it be possible to increase the "score" of a potential association between a gene gap g2X and a reaction gap r4 in M2, given that a gene g1D has a Known association for r4 in M1, even though g2x and g1D share no detectable sequence similarity? Despite the lack of sequence similarity, the contextual evidence might convince a bioanalyst to consider gene g2x as a member of a yet-unknown family able to catalyse r4 (or perhaps some variant of it).

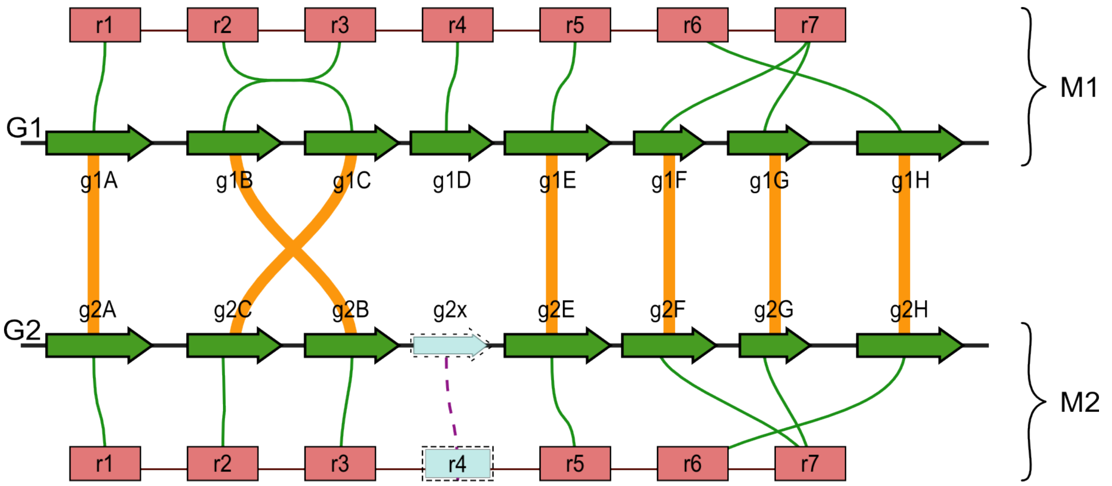


Illustration VII.26: Annotation Transfer by Context Similarity

Two metabolons in two related genomes G1 (genes g1A to g1H) and G2 (genes g2A to g2H). Both concern metabolic reactions r1 to r7. Functional annotations are shown as green curves. High sequence similarities are shown as large orange lines. Gene g1D catalyses reaction r4, and gene g2x is a candidate for the same reaction; however, gene g2x has no identifiable sequence similarity to gene g1D. Given this evidence, a bioanalyst might want to consider g2x as a valid candidate for r4, perhaps belonging to a yet-unknown family of enzymes. In an automatic procedure, it would be interesting to be able to reinforce the association score between g2x and r4 on the basis of that existing between g1D and r4, despite the absence of similarity: an annotation transfer on the basis solely of conserved genomic and metabolic contexts.

Actually defining a procedure for carrying out this kind of transfer is, obviously, a major challenge. I can currently imagine two ways to do this. A first possibility would be to use the expertise of several bioanalysts to create a rule-based system, capable of analysing two or more metabolons in respect to gene, family and reaction content, taking into account gene and reaction similarities, in order to make high-level inferences. A second possibility would be to use an approach akin to that of the Genomic Context Similarity presented in the next chapter, section VIII.C.2. This would require defining some sort of numerical measure of context similarity, be it in terms of genomic context or metabolic context, that could "replace" the absent sequence similarity between them.

Either way, this kind of transfer would be particularly useful for metabolons having been affected by xenologuous gene displacement, or promiscuous enzyme recruitment. However, it would probably generate a large number of false positive potential associations (or up-score them erroneously). It would also be of no use in the case of orphan enzymes: indeed, if g1D only had a potential association with r4, then there would be not enough evidence to hypothesise that g1D nor g2x catalyse r4. For all these reasons, this specific avenue was not explored further in this work.

VII.H.4.b. More functional dependence evidence

As has been suggested in the paper, **additional functional dependence clues** could be added to CanOE. The use of the current version of CCCPart imposes the transformation of these clues into binary gene-gene dependency edges. Clues that could be used include high phylogenetic profile similarity, regulation by a same molecule (requires prediction of individual gene regulations), or high expression profile similarity (requires the use of experimental data). Preparing protocols for these might prove difficult to adjust, especially in selecting a value cut-off for generating binary edges. As gene neighbourhood has been shown to be the most informative of genomic context indicators [163], it would also be necessary to check that their addition does indeed increase CanOE coverage and usefulness. The main expected benefit would be the possibility of having metabolons span multiple locations in a genome, functionally dependent in a way that obviously goes beyond chromosomal proximity. It should thus be possible to capture higher-level metabolons, increasing opportunities for the generation of hypothetical gene-reaction associations.

VII.H.4.c. Non-metabolic genes and functional subsystems

Another kind of improvement for the CanOE strategy (that would pair well with the previous) would be the **inclusion of non-metabolic genes** into metabolons. More specifically, it would be desirable to include non-metabolic genes whose products still participate, in one way or another, in metabolism. This brings the metabolon closer to the notion of "**subsystem**" which is central in the SEED platform [39], and becoming it for MetaCyc [214]. For example, an urea degradation metabolon might benefit from the addition of an urea transporter that happens to be included in the same operon. Other examples might be the inclusion of genes responsible for regulatory processes, signalling cascades, the construction of protein chaperones, *etc*. This might be useful for capturing additional candidate genes, and for making metabolons more representative of the organism's working metabolism. The main hurdle to be overcome is obviously encoding the participation of a given non-metabolic gene in a given metabolic pathway. The exact functions of non-metabolic genes are rarely well described, and not encoded as easily as metabolic activities. Text searching through the MicroScope database for precise descriptions would probably not work. It might be possible to associate genes with specific compounds (not activities) in this way, though how to add these genes to a metabolic pathway remains an open question.

VII.H.4.d. Gene families, similarity matrices

The use of **gene families** is sometimes disputed in the world of bioinformatics, due to the risks in erroneous annotation transfer [53,266]. Indeed, not only is the definition of "gene/protein family" subject to many variations (some families are defined using whole-sequence similarity, others detect functional conserved domains, see section IV.B.2), but actually establishing function on the basis of families is not direct and depends on the former (due to multiple-domain proteins, functional promiscuity, see section V.C.1). It may thus seem preferable to replace the OrthoMCL-based family definition by something else. It might even seem feasible to avoid defining families altogether. In such as case, a more domain-based approach could be selected, akin to EFFICAz [267] or Pfam. Another possibility would be a more context-based approach, such as using IsoRank [173] with protein-protein interaction networks when available.

If one still wishes for the simplicity of whole sequence similarity, then it might at least be interesting to conserve all similarity information, rather than lose some by cutting the sequence universe into non-overlapping families. At one point, I thought that it might be possible to directly use a **similarity matrix** H of size G*G (where G is the total number of genes) of gene-gene similarity scores (*i.e.* the matrix fed to the MCL algorithm). If gene-reaction association scores were to be consigned in an association matrix A of size G*R (where R is the total number of reactions), than a simple matrix product H*A would lead to "association diffusion" amongst the genes. Put simply, a gene would inherit associations to reactions owned by its neighbours, factored by its similarity with said neighbours. Several "rounds" of such multiplication (diffusion) would ensure associations would be shared by all genes with some sequence similarity, and correct associations would "accumulate" in clusters of genes. This would also allow genes to be associated to many reactions, with varying weights. Several pitfalls line this path, however. Firstly, it would be necessary to define prior association values for "Known" and "Potential" associations, as is currently the case, though these choices might affect results more profoundly. Secondly, some sort of normalisation method would be required to actually locate final scores that are indicative of a true association, a feat that is likely to be difficult as maximal association scores will be heavily influenced by localised clusters and clique structures. Finally, CanOE currently generates metabolons for over 200,000 genes. The corresponding H matrix, thought very sparse, would still be difficult to handle computationally. Even the A matrix would be. This lead still seems promising to us, and if further research is encouraging, will probably supersede the current family-based approach.

VII.H.4.e. Aligning multiple graphs at once

As has already been said, the CCCPart implementation developed by Yves-Pol DENIELOU has many algorithmic improvements over the previous version. One of the main new features is the computational tractability of **aligning multiple input graphs in one go** (other improvements include imposing node order constraints, or conservation quotas across multiple input graphs). It could, for example, be used to calculate conserved syntenies across several genomes at once. This could be exploited in order to establish metabolons that are conserved across several genomes at once. The added value of this would be of being able to skip the "integration over n organisms" step entirely, as it is already done. However, several hurdles would have to be overcome before doing this. First of all, the algorithm's output would be highly complex. Indeed, CCCs would be reported across subsets of input graphs (metabolons conserved across variable numbers of genomes), and they would certainly be overlapping (*e.g.* a three-genome metabolon could overlap with a larger two-genome metabolon). This would make dealing with the output very complicated. Secondly, the method would be sensitive to phylogenetic bias, in that heavily conserved metabolons would easily be found amongst groups of closely related genomes; defining genome sets on which to run it would probably have to be done manually. Once again, we stress that the CCCPart algorithm is not adapted to using weighted edges, be it within an input graph or between input graphs; this would be a problem for the gene-gene similarities, as they would have to be made binary, thus losing information, especially since determining a global similarity threshold across the whole gene space would not be biologically relevant. It might be possible to use other algorithms for this task, such as IsoRank or IsoRank-N [173,174], though they were not designed with different types of input graph (*e.g.* genes *and* reactions for metabolons) in mind. Finally, even if integration over n organisms is built in *per se* into the metabolons, this does not determine how individual gene-reaction associations would be scored: an entirely new procedure would be required.

Another possibility of dropping gene family use from the CanOE strategy would be to integrate not genes via families, but entire metabolons. Ways of **aligning metabolon graphs** by using algorithms such as IsoRank [173], [144], or even CCCPart itself [5], or establishing tailored measures of "metabolon similarity" (*e.g.* counting the number of shared genes or reactions, factoring in gene similarities, *etc*) could be imagined to regroup metabolons across many genomes. This would be interesting for comparative genomics, though the same problem of redefining a scoring system for this approach remains.

Other improvements can be imagined, in the way that CanOE data is exploited.

VII.H.5. Metabolon & CanOE use cases

Metabolons are a formalisation of high-level annotation information. Several uses for them could potentially be explored.

The first that springs to mind is their use in predicting **operons**. In their current shape, metabolons are indications of functional coherence in one genome that is in line with the definition of an operon. Perhaps crossing predicted operons and metabolons could help extend the latter. Furthermore, as said in section IV.D.2.c, some operon-predicting algorithms require function to be conserved across several genomes, as well as sequence; multi-genome metabolons (or alignments of several single-genome metabolons) could capture this kind of signal. Perhaps a simplification of this idea would be the use of multi-genome metabolons with loose constraints on its metabolic part in order to derive multi-genome **syntenies**.

Another interesting possibility would be to examine if metabolons associate with other context-gathering units of function (gene runs, gene clusters, syntenies, operons, regulons...). If this was the case, it might be possible to **learn** more relaxed rules of defining **metabolons**, that would in turn be useful for predicting new metabolons, or extending the previous. exploring this lead would obviously be complex and time consuming and could form the basis of a brand new thesis project.

Finally, metabolons could be used to **correct functional annotations**. In [268], the authors adapt their orphan enzymes candidate gene finder to a different problem: that of locating potential mis-annotations. Their idea is that if a gene G is annotated with function belonging to bio-process P, but has low functional dependency with other genes whose annotated functions participate in P, then the annotation of G is probably false. Furthermore, if G has a higher functional dependency with another set of genes than the one its annotation links it to, then the annotation of G is probably false as well. In the latter case, a new function for G can be proposed; this idea is illustrated in the figure below. Metabolons could be used similarly to this. Any gene gap that belongs to a metabolon and whose annotated product does not participate in the metabolon can be viewed as a potential mis-annotation. Our integration over n organisms approach can help, as annotations that are not supported by a metabolon are not considered; a bad annotation that has been transferred by simple homology will thus be thwarted. Potential associations across the family, however, can propose new metabolic functions for it.

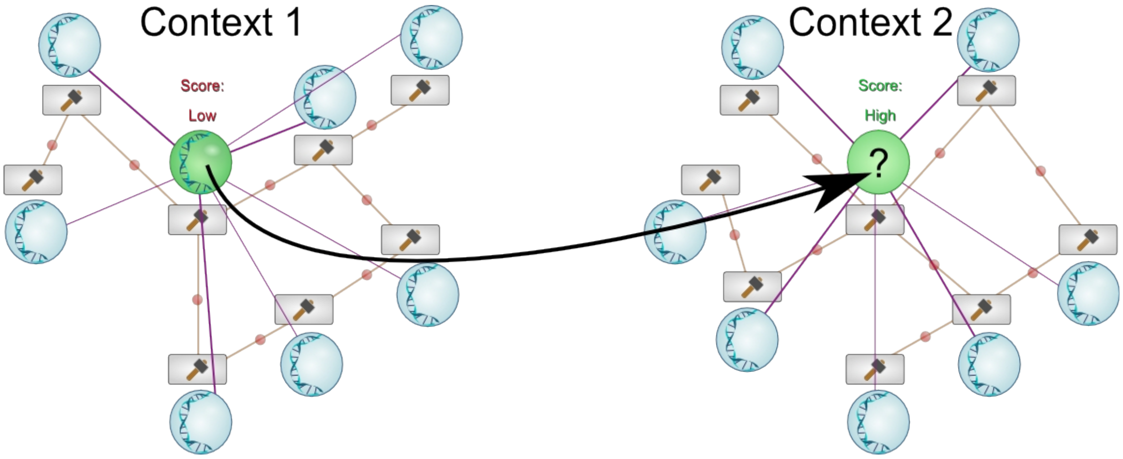


Illustration VII.27: Annotation correction using context-based methods

The figure represents genes from a genome (green and blue circles) and their respective annotated functions (grey rectangles), as well as the links between the latter (orange lines with dots in the centre). In functional Context 1, the gene in green is annotated with one function, but has a low "goodness-of-fit score" for this position, as calculated from its functional dependencies with the other genes (purple lines). However, the same gene has a higher score for a different function in Context 2. The green gene can thus be considered as mis-annotated, and in this case, could be re-annotated with the function from Context 2.

VII.H.6. Enzymatic promiscuity

As I have already pointed out, I believe any new bioinformatics method dealing with functional annotation should in some way deal with functional promiscuity. Indeed, as has already been highlighted out in [97,98,100], the relationship between sequence similarity - the most heavily-used proxy for functional similarity - and function is far from straightforward, and functional promiscuity is one of the mechanisms that clouds it up.

The first version of CanOE does not deal with promiscuity either. Worse, some of its working hypotheses expect genes to be mono-functional (*e.g.*, not proposing already annotated genes as candidates for a reaction gap). Still, one might excuse CanOE, given that I was unable to find anything in the literature that approached this problem. We have, however, imagined some ways of dealing with functional promiscuity in the CanOE framework, and I wish to present them here on the off chance that they may one day be of some use to future developers.

It has already been question about the "**underground metabolism**" [230], the idea that the genes encoding one metabolic pathway may actually be capable of catalysing parallel or even completely different metabolic pathways, either at negligible speeds, or only when the right conditions are met (such as the absence of the first substrate of the primary pathway). The CanOE metabolons can capture some of this, either by a) allowing genes to be associated to multiple reactions in a same metabolon, or b) allowing genes to be assigned to multiple metabolons. Case a) would capture parallel or similar pathways whose reactions share, at some point or other, identical main compounds. Case b) can capture different pathways or those who do not share main compounds. Cases like these were observed within CanOE, especially for KEGG reaction-based metabolic schemas, due to the uncertainty of EC to KEGG reaction conversions (often allowing one gene to associate with a panoply of related KEGG reactions). However, most of these cases seemed anecdotal or spurious, and depended on genes that have already been annotated as multi-functional. Underground metabolism is currently expected to be largely unknown, so the CanOE strategy as is may be useful in filling in blanks, but will probably not help discover these cases directly. To be able to live up to this mission, CanOE would have to be improved in some manner beforehand.

The main point to address prior to improving CanOE in this direction is "how exactly to **model functional promiscuity**?". The current model allows for it by accepting multiple Known gene-reaction associations for a same gene. Additionally allowing Potential associations to be generated for genes with Known associations would further open CanOE up to functional promiscuity, but would result in creating many, many false positive Potenial associations. One interesting alternative would be to model reaction-reaction similarities in a graph. These similarities would be designed to capture enzymatic promiscuity. For example, a reaction R could be proposed in a Potential association for a gene G that already has a Known association with reaction R', if and only if R and R' are two enzymatic reactions that can be catalysed by a same enzyme. More boldly, low-confidence Known associations could be automatically generated on the basis of these similarities, thus hypothesising that if an enzyme catalyses one reaction that is known to be catalysed with another, then it probably also catalyses the other.

The simplest way that springs to mind of constructing such a similarity is to count the number of **co-annotations** (*i.e.* the number of times a gene has been annotated with reaction R and reaction R') and to divide it by the number of individual annotations (*i.e.* the number of times a gene has been annotated with R or with R'). This approach would ensure that reactions that are known to be able to cohabitate within a same enzyme are effectively given this possibility when designing novel Potential associations. Another possibility would be to base reaction similarities on **EC number sharing**. This is implicitly the case in [244], where candidate genes are favoured when they share the first three digits of the EC number corresponding to the target gap reaction. However, the power of the EC number system to capture reaction similarity has been disputed, either because of the accumulation of example cases where EC number similarity could not be traced to sequence similarities [269,100], or because of discrepancies with reaction mechanistic knowledge [270].

Reaction similarity should thus take the chemical reaction into account, preferably with knowledge of mechanistic steps. [271] built reaction-describing feature vectors based on reactant structural physico-chemical properties observed to change during the reaction. They classified these features using factorial analysis or Kohonen self-organising maps and derived a reaction similarity measure from the resulting spaces. In [270,199], the authors propose a more mechanistically-orientated approach: a) individual mechanism step similarity between reactions was evaluated on physico-chemical properties, transformations or bond changes, b) aligning the steps of 2 reactions using step-to-step similarities using the Needleman-Wunsch algorithm, and c) deriving a normalised score from the alignment to represent reaction similarity. The drawback of mechanistically-precise approaches is that they are obviously limited to reactions for which the mechanism has been elucidated. The latter are far and few between, as state-of-the-art databases dedicated to this sort of information, such as the MACiE database [272], only cover 321 EC numbers (for a total of 335 distinct reaction mechanisms) at the time of writing [www.ebi.ac.uk/thornton-srv/databases/MACiE].

Work conducted by Syed Asad Rahman at the Thornton Group of the EBI recently presented at ISMB/ECCB 2011 dealt with deriving reaction similarities on the basis of **chemical structure transformations** between substrates and products, a biologically-relevant approach that does not depend on reaction mechanism. However, these works are still in progress, though the first publications should be coming out soon. Once they become available, they should provide another way of measuring reaction similarity that is adapted to enzymatic promiscuity.

Once all similarities between reactions have been obtained, it remains an open question on **how to use them**. The methodologies hinted above propose to use some sort of binary rule, allowing or disallowing the generation of Potential associations on the basis of reaction similarity. Another possibility involves using the matrix-based integration approach discussed in section VII.H.4.

Reaction similarities would be put into a matrix T of size R*R (R being the number of reactions). Diffusing gene-reaction association scores across matrix A on the basis of reaction similarity could then be done simply by taken the matrix product A*R. Iteratively setting $A_n = H * A_{n-1} * R$ would ensure that association information would be diffused according to both gene and reaction similarities. It would arguably be even more complicated, however, to devise a normalisation scheme to extract conclusions out of a resulting A_n matrix.

All these leads and ideas are potential starting points for future improvements to the CanOE strategy, and it is my fond hope that someday they may help continue to carry research forwards in the domain of prokaryote genome annotation.

VII.I. Strategy conclusion

The CanOE strategy is a useful contribution to the domain of automatic gene function prediction for several reasons:

- it generates metabolons, a useful unit of metabolic function in a genome, that can be readily exploited by bioanalysts to ease manual annotation;
- it can propose potential and inferred associations and between un-annotated genes and reaction gaps, with scores integrated across multiple genomes, thus helping bioanalysts to focus on the most promising cases;
- it can even propose candidate genes for organism-orphan reactions (albeit in small numbers) without use of sequence similarity, which is of particular interest for global/prokaryote orphan reactions;
- it uses a computationally efficient and exact algorithm (with biological priors to result post-processing) to locate genomic metabolons in single prokaryote genomes;
- though it exploits metabolic context and solely gene co-localisation as genomic context information, it should be adaptable to other types of useful contextual information, such as phylogenetic profiles and co-regulation (this will probably form the basis of a new doctoral or post-doctoral project at the LABGeM);
- unlike competitor methods, CanOE is the first - to our knowledge - to propose the computational integration of results across all available genomes, leading to the calculation

of family-reaction association scores that we showed to be informative for bioanalysts in prioritising gene candidates for target reactions;

- competitor methods such as [144,244] do not propose their results to the scientific community via a web interface or downloadable data sets, or even downloadable programs or source code. Results of the CanOE strategy will be fully integrated into the MicroScope web-based prokaryote genome annotation platform (though source code, presenting no particular general interest and being MicroScope-specific, will not be made available).
- the CanOE strategy shall be presented to the scientific community via two scientific publications, one detailing the strategy, inserted above, and another describing the latest MicroScope developments. The MicroScope platform now offers users the possibility to directly annotate genes with MetaCyc reactions (as well as previously-used EC numbers), a development which should form an interesting synergy with CanOE.
- The CanOE strategy provides interesting development and use case perspectives that would be well worth exploring in the future.

VIII. Development: The BKACE project

The strategy imagined and implemented here will be presented in a collaborative paper regrouping the many Genoscope actors of the entire project, wherein the bioinformatics part will be relatively restricted. Full details of my work are thus discussed here.

VIII.A. Motivation and Objectives

In 2007, a collaborative work between the LABGeM and the LGBM teams at the Genoscope led to the discovery of the coding genes for three reaction steps in the lysine fermentation pathway that had previously evaded identification [240]. A comparative genomics approach using the MicroScope platform showed that a group of co-localised genes were conserved across several genomes, several of which corresponded to genes encoding enzymes from the then-incomplete pathway, the others encoding hypothetical proteins. With this strong evidence of functional association, wet-lab experiments were carried out in order to elucidate the functions of the un-annotated genes, successfully assigning the three gene-less steps to members of three hypothetical gene families.

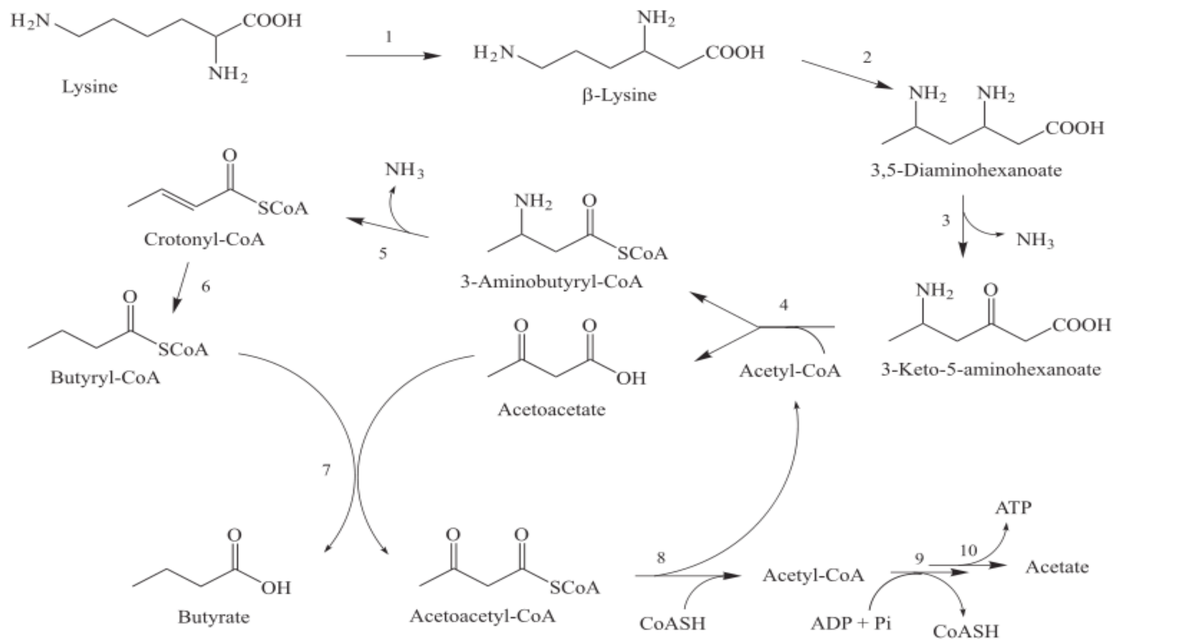


Illustration VIII.1: The lysine fermentation pathway

Shown enzymatic reactions are: L-lysine-2,3-aminomutase (1), beta-L-lysine-5,6-aminomutase (2), 3,5-diaminohexanoate dehydrogenase (3), 3-keto-5-aminohexanoate cleavage enzyme (**BKACE activity**) (4), 3-aminobutyryl-CoA ammonia lyase (5), acyl-CoA dehydrogenase (6), acetoacetate:butyrate CoA-transferase (7), acetoacetyl-CoA thiolase (8), phosphate acetyltransferase (9), and acetate kinase (10)

One of the involved gene families was classified in Pfam [76] as “DUF849”, meaning that the family had been built on the presence of a conserved protein Domain of Unknown Function, and several members of this family were shown to catalyse the 3-keto-6-aminohexanoate cleavage step, illustrated in [Illustration VIII.1: The lysine fermentation pathway] (figure borrowed from [240]). However, this function was not shared by ALL the members of this family, of which most came from organisms unable to ferment lysine. This and the high conservation of the protein domain suggested that proteins from this family were able to catalyse a broad range of enzymatic activities, affecting different substrates by a similar chemical mechanism: enzymatic promiscuity at least at the family level (though promiscuity at protein level, *i.e.* a same protein being able to catalyse several reactions, was not ruled out). The project was then extended, with the objective of a) characterising the 3D structures of representative proteins of this family, b) elucidating the associated mechanism, and c) exploring the functional diversity of the family. The first and second points were accomplished by a collaboration between the Structural Microbiology team at Institut Pasteur and two other Genoscope teams: the Laboratoire de Chimie Organique et Biocatalyse (LCOB) and the Laboratoire d’analyses bioinformatiques des séquences (LABIS) [273][19]. Protein 3D structure modelling carried out in the LABIS showed that the active site of the enzyme was very probably conserved across the family. Together with the proposed mechanism, this comforted the hypothesis of a set of similar activities acting on similar substrates, and led us to name the family “BKACE”, for “beta-keto acid cleaving enzyme family”.

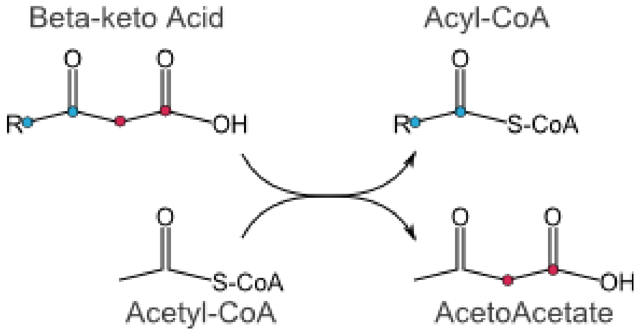


Illustration VIII.2: The generic beta-keto acid cleavage reaction

Substrates are a beta-keto acid and an acetyl-CoA molecule. An acetyl group (carbons in red) is transferred to the acetyl-CoA's acetyl group, generating acetoacetate, while the CoA replaces it in the beta-keto acid, giving an acyl-CoA (conserved carbon chain in blue). We hypothesise that the variable part of the beta-keto acid (R) can contain many different functional groups: alkyl chains, alcohols, amines, unsaturated alkyl chains...

[19] Though this study was limited to a single lysine degradation-related BKACE protein.

My work in this project addressed two main points of the exploration of the functional diversity of DUF849. The first of these concerned using existing bioinformatics tools to help divide DUF849 into subfamilies that would hopefully prove to be iso-functional (*i.e.* act upon a same set of similar substrates). This would reduce the protein space to explore from all 749 proteins from DUF849 to only several hundred (choosing, for example, 192 that would conveniently fit into two 96-well plates during wet-lab experiments). Furthermore, using bioinformatics tools, chemoinformatics tools and expert knowledge, we wanted to propose a set of plausible substrates that the proteins could act on.

The second point I intervened on was the statistical analysis of the experimental data that was finally obtained by the LCAB team.

VIII.B. Article

A collaborative paper is currently being written in order to publish our findings. Its advancement, however, does not allow me to insert the draft here. Instead, I shall insert the short submission I sent in order to present these works at the "Journées Ouvertes de la Bioinformatique et des Mathématiques" (JOBIM) conference in September 2010. The submission was accepted for a 5-minute flash oral presentation and a poster.

Author Contributions and Acknowledgements: *Many different teams collaborated on this project: the Laboratoire de Génomique et de Biochimie du Métabolisme (LGBM), the Laboratoire d'Analyses Bioinformatiques de Génomique et du Métabolisme (LABGeM), the Structural Microbiology team, the Laboratoire de Chimie Organique et Biocatalyse (LCOB), the Laboratoire d'Analyses BIoinformatiques des Séquences (LABIS), and the Laboratoire de Clonage et de criblage des Activités de Bioconversion (LCAB).One can only acknowledge as a whole the work provided by all the team members that were involved in the BKACE project.*

VIII.C. Methods and results

Dividing a protein family into subfamilies supposes the use of at least one type of protein similarity measure to calculate an optimal partition. Following the general trend in bioinformatics, our approach used several protein similarity data sources and integrated them together, hopefully reaching increased method precision and coverage. Here, I shall present the methodologies for each individual similarity, as well as the integration method. The complete clustering pipeline is represented in the figure below. I shall then detail the experimental protocol used, as well as the statistical treatments I applied to its biochemical results, insisting on some points of the experimental design that motivated my choices.

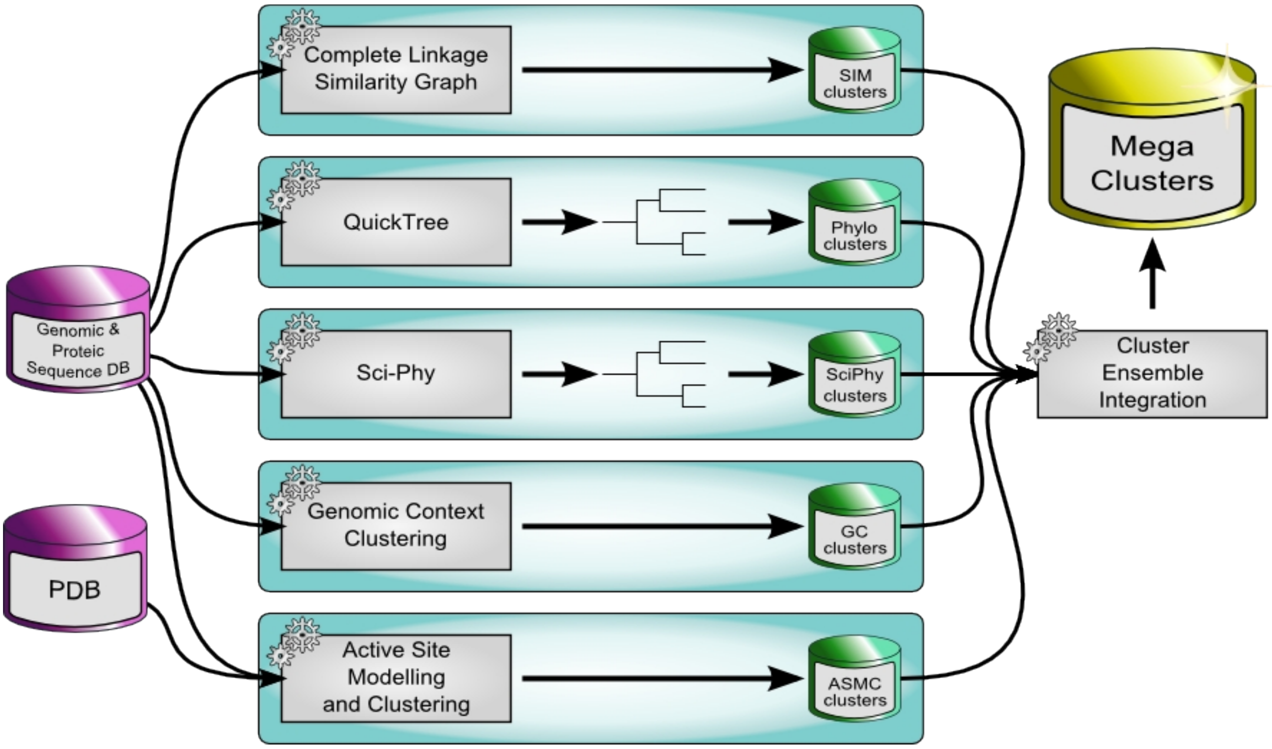

Complete Linkage Similarity Graph
SIM clusters
QuickTree
Phylo clusters
Sci-Phy
SciPhy clusters
Genomic Context Clustering
GC clusters
Active Site Modelling and Clustering
ASMC clusters
Genomic & Proteic Sequence DB
PDB
Cluster Ensemble Integration
Mega Clusters

Illustration VIII.3: The BKACE clustering pipeline

Primary sequence and structure data are extracted from public databases. Five main component BKACE clusterings are carried out: complete linkage, quicktree, Sci-Phy, Genomic Context, and ASMC clusterings. These are then integrated using the cluster ensemble framework into a single “MegaClustering”, the clusters of which are supposedly iso-functional. The Genomic Clustering is also used as a basis for finding coherent metabolic contexts from which to draw potential BKACE substrates.

VIII.C.1. Multiple Sequence Alignment

The study of gene or protein families usually starts with the examination of all involved nucleic or amino acid sequences. Tools such as BLAST determine optimal pairwise sequence (nucleotide or amino-acid) alignments. However, to build representations for several proteins and to study sequence evolution, it is necessary to align several proteins at once, making what is called **Multiple Sequence Alignments** (MSAs). Specific algorithmic challenges arise for this sort of problem and several tools have been created in answer, such as Clustal [274], MUSCLE [275], and the very recent Clustal OMEGA [276]. In this work, we used MAFFT [277]. MSAs can be used to establish phylogenies of the involved sequences by establishing protein-protein similarities that benefit from the family alignment, which from an evolutive point of view should be more precise than simple 1-*versus*-1 sequence similarity. In practice, they can also be used to filter out proteins that do not align very well with the rest, such as proteins with sequencing errors or fragmented proteins. Here, 739 BKACE proteins were initially aligned, but we filtered them down to the 725 that will be used in the subsequent analyses.

VIII.C.2. Clustering on conserved genomic context

We used an in-lab tool called the **Syntonizer** (developed by Laurent LABARRE and David VALLENET [6]) to locate syntenies (according to the MicroScope definition given at the end of section IV.D.2.c). It uses the CCCPart algorithm from [5] (see section VII.C.1), using edges based on discrete gene contiguity and binary gene homology relationships. The Syntonizer was used in establishing genomic context similarity measures presented in the following section.

We developed an original clustering approach specifically for the BKACE project. The general idea was to cluster together BKACE proteins that had similar genomic contexts. We decided to build a **measure of genomic context similarity** based on the **conservation of gene neighbourhoods** between BKACEs identified using the previously described Syntonizer. I implemented the following protocol :

- For each and every BKACE-coding gene, extract from MicroScope all the neighbouring genes that are entirely within a 10,000 base-pair distance (the average CDS length in *E. coli* K12 is 960 base pairs, and the average intergenic distance is 142 base pairs, meaning we can capture on average 9 genes on either side of the BKACE gene). This will be referred to as a

BKACE neighbourhood.

- For each concerned organism pair (*nota bene*: several BKACEs can exist in a same organism, so self-pairs are allowed), extract all gene-to-gene sequence similarities as calculated by BLAST for all genes in all neighbourhoods (including the BKACE proteins themselves).
- Find all syntenies (*i.e.* groups of conserved genes) between all BKACE neighbourhoods using the Syntonizer with the following parameters:
    - ○ Gene similarity was established using BLAST alignments of the sequences of the encoded proteins. Homology was concluded between two best-bidirectional best hitting proteins that had at least 30% amino-acid identity over 80% of the length of the shortest protein[20], which are the default settings for syntenies determined in MicroScope [6].
    - ○ Syntonizer gene gap parameter was set to 3, which is more restrictive than MicroScope base settings (which allow 5 gap genes), and only syntenies involving more than 3 gene-to-gene relationships were conserved.
- Drop all syntenies that do not include a BKACE protein as they probably don't concern the latter.
- Using each remaining synteny (thus linking a BKACE and part of its gene neighbourhood on one organism to another BKACE and its neighbourhood from the same or another organism), establish a genomic context similarity measure between the involved BKACE proteins.

We decided to choose an intuitive and simple measure of genomic context similarity: **the number of genes shared between two neighbourhoods in a synteny**. It is intuitive that the more genes conserved between two organisms, the more the BKACE contexts should be similar, and hence the more their functions are probably similar. This approach might be improved by taking into account the actual similarity scores involved.

BLAST results are inherently **asymmetrical** due to the nature of the algorithm, *i.e.* the similarity of protein A to protein B is not necessarily identical to the similarity of B to A. Furthermore, syntenies are not necessarily symmetrical either, *i.e.* the number of conserved genes of neighbourhood A in

[20] *Nota bene*: please note that not all pairs of BKACE proteins actually pass these criteria. Thus, two heavily dissimilar BKACE proteins cannot be part of a same Genomic Context in our protocol.

neighbourhood B is not necessarily the same as the number of conserved genes of B in A, as there can be gene fusion/fission and duplication events. Four possible counts of conserved genes are thus available for each pair of BKACE proteins. To establish a single measure, we thus chose the minimum number of conserved genes from each version of a same synteny (neighbourhood A *versus* neighbourhood B and B *versus* A) and summed the values for each version of BLAST (organism A *versus* organism B and B *versus* A), which is equivalent to taking the average. This is illustrated in the figure below.

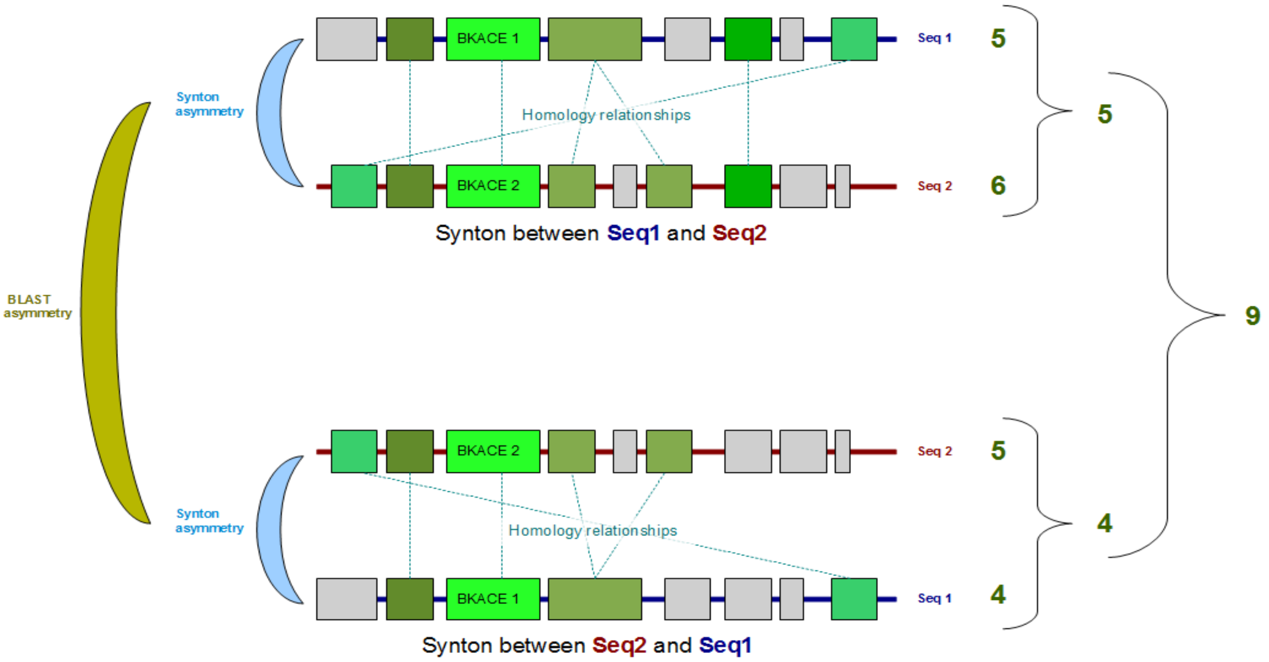
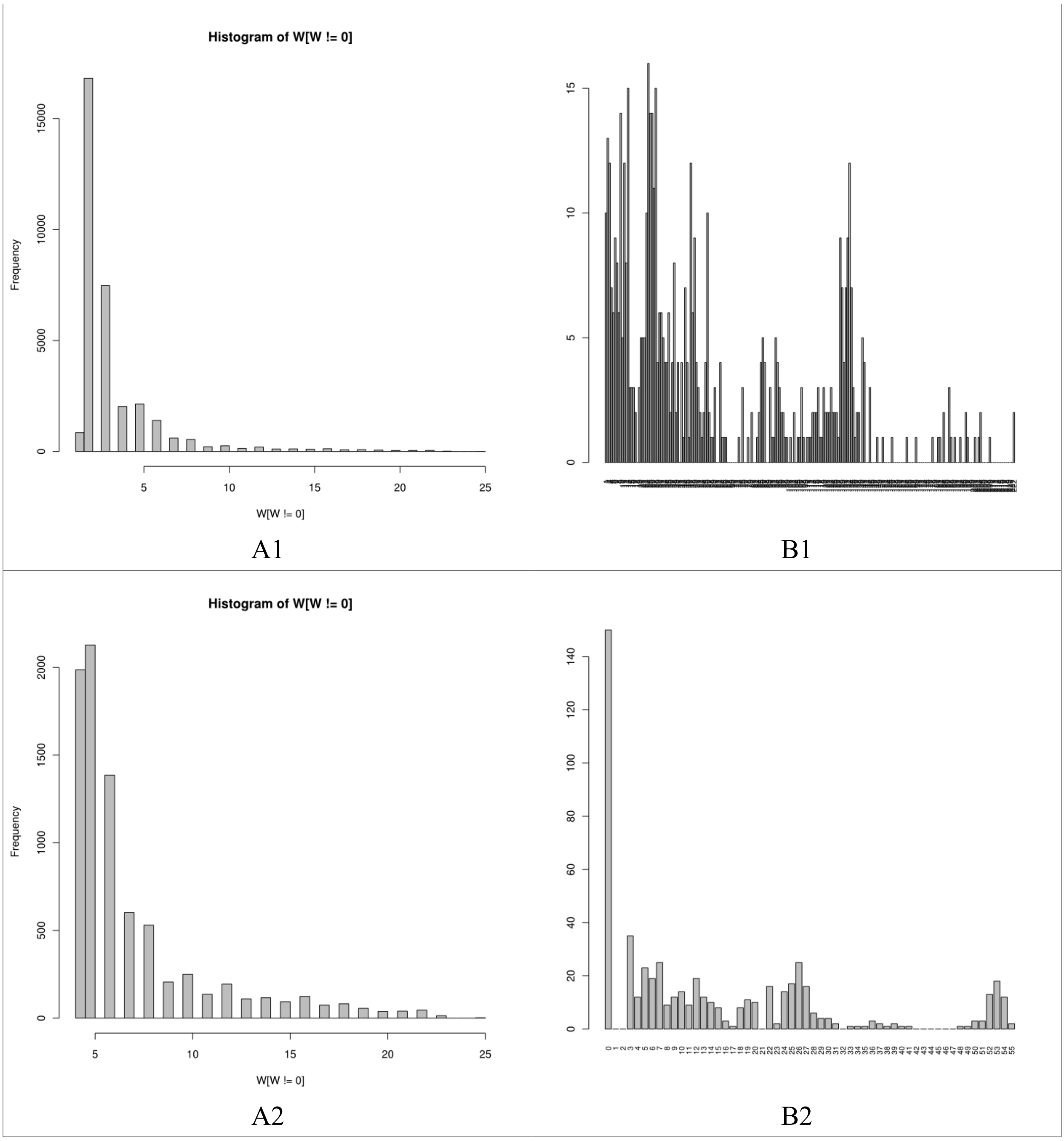


Illustration VIII.4: The Genomic Context similarity measure

Genes from a synteny between genomic sequences Seq1 and Seq2 are shown in green, with BKACE proteins in striking green. The Genomic Context Similarity measure is an approximation of the number of genes in synteny. The asymmetry of synteny results (due to fusions/fissions) is dealt with by taking the minimum number of involved genes from one genome. BLAST result asymmetry is dealt with by taking the sum of the previous for each possible setup (Seq1 *versus* Seq2, and Seq2 *versus* Seq1).

The resulting counts were used to weight edges in a graph, wherein each node represents a BKACE gene neighbourhood. This graph required **pre-processing** in order to make its latent structure more apparent, and to render it more robust to perturbations. Indeed, this kind of similarity graph is not transitive: if neighbourhood A is similar to neighbourhood B and B to C, A is not necessarily similar to C, or can be similar to C, but via a different set of genes. To address these shortcomings, we iteratively removed edges that were a) too weak (minimum edge weight:4) and b) connected nodes with too low a degree (minimum node degree:3), ensuring minimal levels of connection strength and connectivity, until convergence of graph structure. Distributions are given in the following figures for illustrative purposes:



A1

B1

A2

B2

Illustration VIII.5: Edge weight and vertex degree distributions in the Genomic Context graph

The distributions of edge weights (A) and vertex degrees (B) is shown for the Genomic Context graph before (1) and after (2) the iterative filtering.

We then applied a recursive **spectral clustering algorithm** (a class of clustering algorithm to which the tutorial in [278] can be considered a practical introduction) to divide the graph into dense modules with sparse inter-connexions. These yielded the BKACE clustering based on genomic context similarity that we desired, called “**GC clustering**”. Unfortunately, due to the pre-processing steps, many BKACEs were singletons in the graph, and could thus not be clustered. The number of clusters was manually fixed to 32, covering 412 BKACE proteins (56.8% of all BKACE proteins). The Genomic Context graph is shown in the following figure, with vertices coloured according to cluster membership.

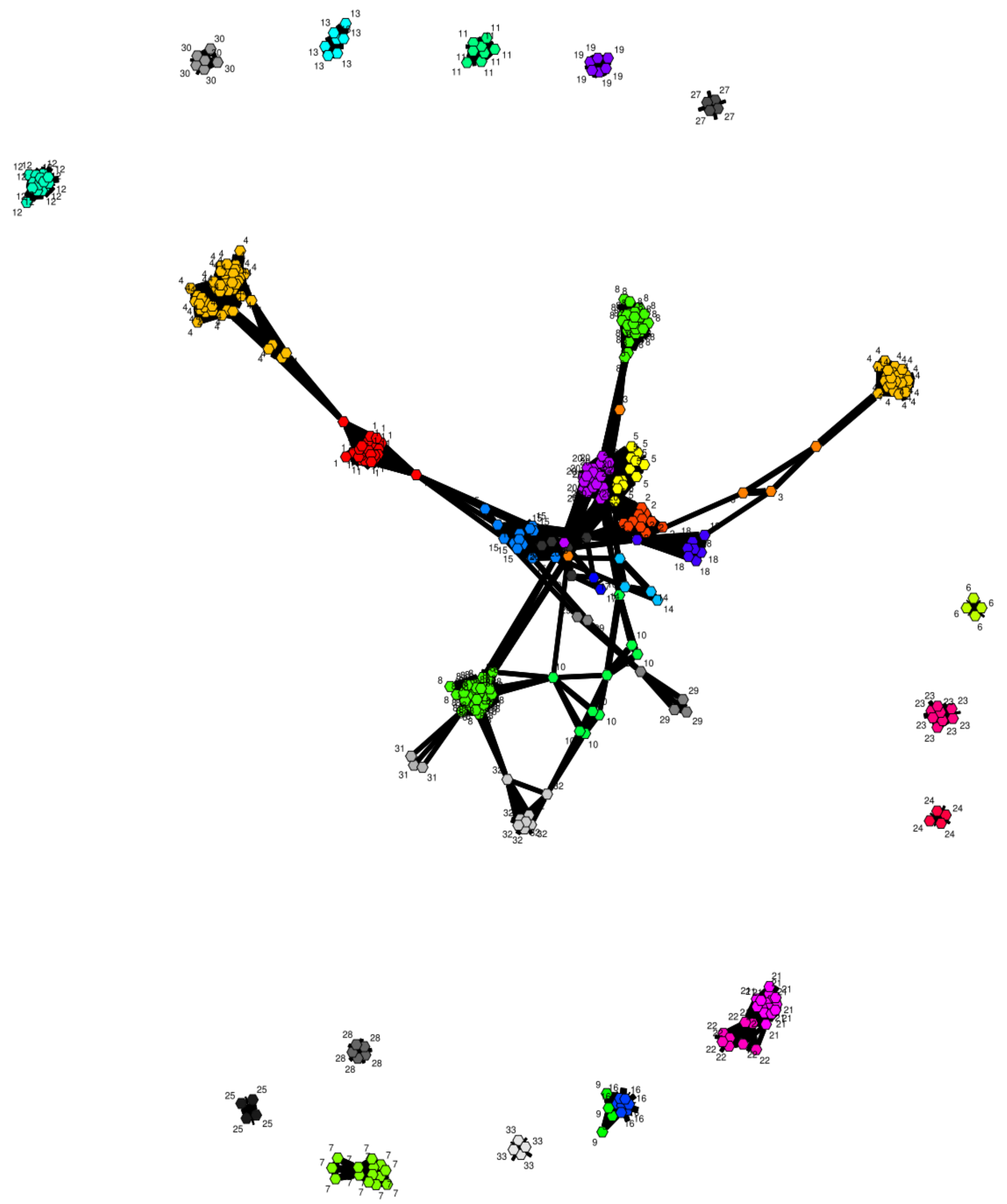


Illustration VIII.6: Genomic Context graph with clusters

Each vertex in the graph corresponds to one BKACE protein and its neighbouring genomic context. Each edge represents a similarity, the score of which passed the pre-processing filters. Vertices are coloured according to the final Genomic Context clusters obtained by spectral clustering of the graph.

VIII.C.3. Clustering on sequence similarity

Clustering proteins on the basis of sequence similarity has been heavily used over the years. Using 1 *versus* 1 protein-protein similarities calculated by BLAST alignments, one can represent the proteins in a similarity graph to which traditional graph-clustering algorithms can be applied. We filtered the BLAST results using manually-established thresholds on several criteria after observing their distributions in the data: matches covering less than 75% of the length of the shortest protein were dropped; matches less than 200 base pairs were dropped (the conserved domain on which DUF849 was built is roughly 200 base pairs long); matches with an e-value above 10^{-50} were dropped. Gene similarity was then summarised by a common score: the opposite log10 of the BLAST e-values.

The similarity graph S was made symmetrical (by averaging scores of A against B and B against A) and transformed into a distance matrix D using the usual formula: $D_{x,y} = (Max(S) - S_{x,y}) / Max(S)$. We then built a hierarchical tree of BKACE proteins from this distance matrix using the base complete linkage algorithm available in the R statistical scripting language. Clusters were finally determined traditionally, using a manually-chosen (after observing the tree structure) height cut-off, which led to 88 clusters of which 49 (representing 6.6% of the 739 BKACE proteins clustered this way) were singletons.

This clustering was called "**SIM clustering**" for simple similarity-based clustering, and had 39 non-singleton clusters covering 690 (95.2%) of all kept BKACEs.

VIII.C.4. Phylogeny-based clustering

We used the DUF849 MSA in two phylogeny-based methods for clustering.

First, we created a phylogenetic tree using **QuickTree** [279], generating bootstrap values for all nodes (which basically gives an indication of the robustness of the tree structure at each node regarding small perturbations in the sequences). I then proceeded to **manually** cut branches of the tree in order to determine the clusters, taking into account tree topology, cluster size and the bootstrap values, which is the traditional approach of the phylogenetician. In this "**PHYLO**" clustering, 71 non-singleton clusters were established, covering 701 BKACEs (96.7% of 725).

Secondly, we used a more advanced and automated approach called **SCI-PHY** (Subfamily Classification in Phylogenomics) [280]. SCI-PHY builds a hierarchical sequence tree using Dirichlet mixture densities to construct sub-tree sequence profiles and using relative entropy as a measure of profile-to-profile or sequence-to-profile similarity. It does not use information on host species phylogeny. It then applies minimum description length principles from information theory to cut the tree into optimal subfamilies. Full details on the statistical protocols can be found in the method's publication and references therein [280]. This clustering generated 46 non-singleton clusters covering 704 BKACEs (97.1% of all BKACEs).

VIII.C.5. Clustering on key amino acids

Within an enzymatic protein, several amino-acids in the sequence are of particular importance. Indeed, beyond the large part of amino acids responsible for the secondary and tertiary structures of the protein, several directly can intervene in the enzymatic activity's mechanism, such as by offering a proton donor group at the right place, or deforming a substrate's electron cloud so as to facilitate group substitutions. Other amino acids are responsible for constraining the molecules involved in the reaction to spatial conformations more favourable to the mechanism. Identifying key amino acids is an active area of research in structural chemoinformatics, as they can help understand chemical mechanisms and their evolution.

In the BKACE project, the (now disbanded) Laboratoire d'Analyses BIoinformatiques des Séquences (LABIS) contributed -amongst other things- by using their in-lab method called **ASMC** (Active Site Modelling and Clustering) to identify important amino acids in the family [95]. The methodology is summarised in [Illustration VIII.7: ASMC pipeline] (as borrowed from [95]), and is described hereafter:

- retrieve set of proteins from a target family.
- retrieve known 3D structures for proteins of the family, or of proteins similar to those of the family, from the **Protein Data Bank** (PDB) [91].
- use the homology modelling tool **MODELLER** [281] to establish 3D models for all proteins.
- use **Fpocket** [282] to detect the putative active site pocket in the known models.
- use **MultiProt** [283] to align all 3D models.
- retrieve spatially- (and **not** sequence-) aligned amino acids from the active site pocket and represent as an **MSA.**
- create a hierarchical tree of proteins based on this MSA using the **CobWeb** algorithm [284]

from the Weka software [285].

- use a **log-likelihood analysis** statistical approach [286] to identify sub-families and key sub-family segregating amino-acid positions within the MSA.

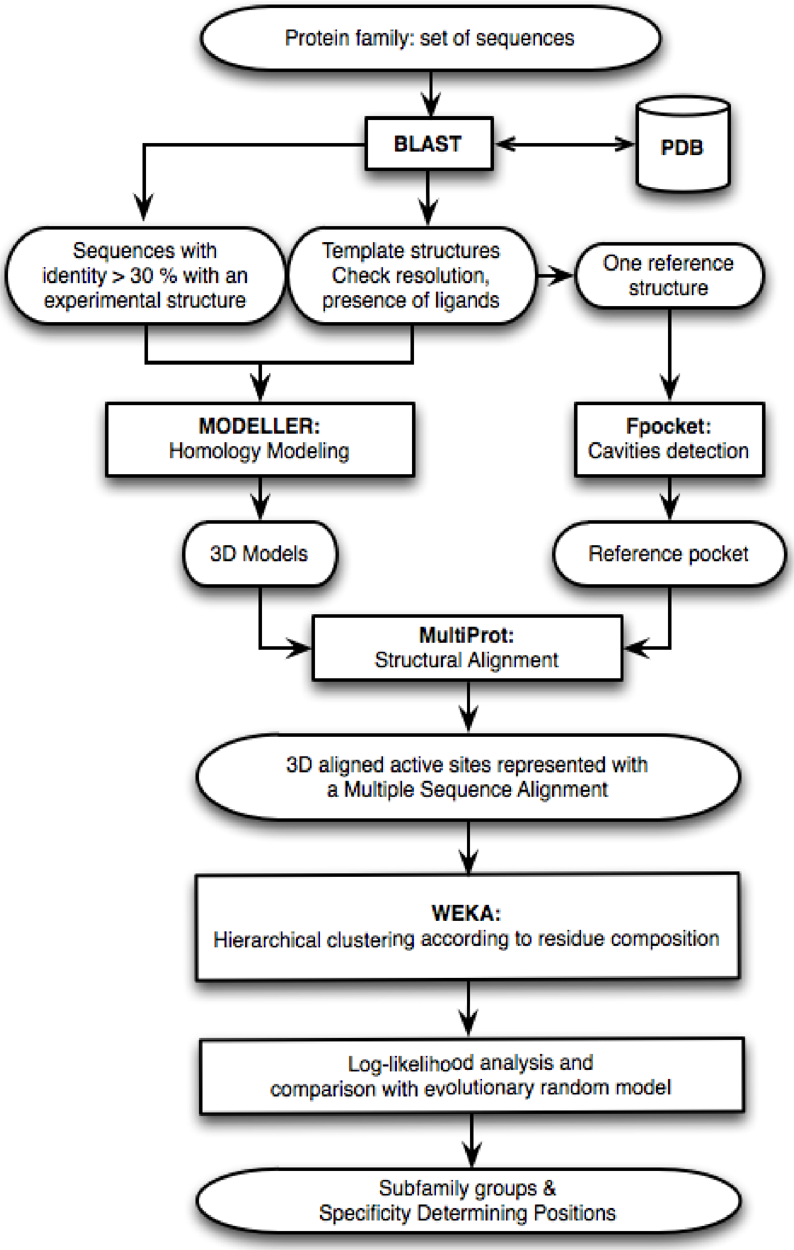

Protein family: set of sequences
BLAST
PDB
Sequences with identity > 30 % with an experimental structure
Template structures Check resolution, presence of ligands
One reference structure
MODELLER: Homology Modeling
Fpocket: Cavities detection
3D Models
Reference pocket
MultiProt: Structural Alignment
3D aligned active sites represented with a Multiple Sequence Alignment
WEKA: Hierarchical clustering according to residue composition
Log-likelihood analysis and comparison with evolutionary random model
Subfamily groups & Specificity Determining Positions

Illustration VIII.7: ASMC pipeline

The bioinformatics pipeline of the Active Site Modelling and Clustering method uses a set of initial protein sequences from a given family, as well as at least one 3D structure for one of the family's members. Cavity detection allows the retrieval of the 3D coordinates of the active site. Homology modelling ensures the retrieval of the amino acids involved in the active site for all proteins. A statistical analysis allows the isolation of key amino acids from the set, and the clustering separates the protein family into subfamilies based on these key residues.

The method results in a protein subfamily clustering as well as knowledge of which amino acids are highly conserved in the active site, and which ones are discriminant between subfamilies. This clustering was called "**ASMC-Cobweb**" and had 31 non-singleton clusters, covering 672 BKACEs (92.7% of 725)

An additional clustering of ASMC active site residues was attempted using multiple correspondence analysis. This clustering was called "**ASMC-MCA**" and had 15 non-singleton clusters, covering all 725 BKACEs. I decided to keep both clusterings for ASMC as each was able to isolate different kinds of extreme cases that the other did not, as shown in a manual analysis by Karine BASTARD.

High-level clustering results are summarised in the following table:

Clustering	Coverage[21]	Non-Singleton Clusters
GC	562 (77.5%)	33
SIM	690 (95.2%)	39
PHYLO	701 (96.7%)	71
SCI-PHY	704 (97.1%)	46
ASMC-Cobweb	672 (92.7%)	31
ASMC-MCA	725 (100%)	15

VIII.C.6. Cluster ensembles

Historically, a first foray into BKACE functional space was conducted on the basis of an *ad hoc* rule-based clustering, integrating results from previous PHYLO, GC and SIM clusterings. This approach was later abandoned due to its lack of rigour, but was useful in confirming in a preliminary biochemical assay that the approach was indeed promising. The framework that superseded it is known as cluster ensembles.

VIII.C.6.a. Protocols

Cluster ensembles is a statistical framework that deals with integrating multiple clusterings/partitions of a same set of objects [287]. It is mostly used to combine multiple component clusterings in a particular way so as to obtain one single integrated "consensus" clustering, and can also establish hierarchical trees of clusterings to help compare them.

Our objective here was to obtain a final protein subfamily clustering based on the individual

[21] Coverage is given only for non-singleton clusters.

clusterings previously discussed. We used **CLUE**, an R package implemented by Hornick *et al.* [288].

Several different choices can be made in CLUE to determine how the integrated clustering is established from the component clusterings:

- component clusterings of a same set of objects can be attributed different **weights** modifying their influence on the integrated clustering.
- the integrated clustering can be calculated using several methods that attempt to optimise a given **summary statistic** that measures how "close" the integrated clustering is to all the component clusterings according to a **clustering similarity metric**.
- the integrated clustering can be either "**hard**" (objects are attributed to a single cluster) or "**soft**" (objects have probability-like attributions to one or several clusters).
- finally, CLUE does not support missing values for object clusterings (*i.e.* unclustered objects in a component clustering). Two approaches are possible for adapting our component clusterings with missing values: 1) as missing values correspond or can be considered to correspond to **singletons**, they can be replaced by artificial singleton clusters 2) as missing values report the inability of the method to cluster the given BKACE proteins, they can be all added to a single "**dustbin**" cluster.

Several tools are available to evaluate the quality of on integrated clustering. The **clustering tree** is a hierarchical classification of the component and integrated clusterings, based on the clustering similarity metric. The **agreement matrix** is an n*n matrix (where n is the number of objects clustered) where each cell (i, j) counts the number of different component clusterings that cluster object i and j together (modulated, if necessary, by component weights). High agreement amongst component clustering leads to many high and low values in the matrix, and few intermediate values. When represented graphically using a colour scheme reflecting count value, if the rows and columns are permuted properly, a block structure should be visible across the matrix diagonal when the agreement is strong. The quality of an integrated clustering can thus be appreciated by visually evaluating this block structure after ordering the rows and columns according to integrated clusters (themselves ordered by cluster size for best visual effect).

I experimented with component cluster weights, cluster integration methods, their parameters, and the singleton/dustbin approach to singleton/missing value management. I also experimented with using a factorial analysis in order to integrate the clusterings (a multiple correspondence analysis, followed by a hierarchical clustering in the resulting factorial space). Resulting integrated clusterings were manually evaluated on the basis of the number of final clusters, the number of final singletons, on the aspect of the agreement matrix, of the clustering tree...

All of the component clusterings used here were born of trees, so it would seem natural to try and establish the tree behind the consensus clustering. Please note, however, that the consensus clustering is not based on a tree, but is the direct result of an optimisation algorithm using clustering metrics. It is thus not possible to obtain a hierarchical tree of the objects clustered in the consensus clustering. In order to derive a consensus tree, one would have to turn to the tree ensembles statistical framework (as noted in the discussion later).

VIII.C.6.b. Results

The agreements between the various component clusterings (in both their dustbin/singleton versions) and the possible consensus clusterings (*i.e.* Factorial analysis, Hard Manhattan/Hard Euclidean, weighted/unweighted versions) are illustrated in the following agreement tree:

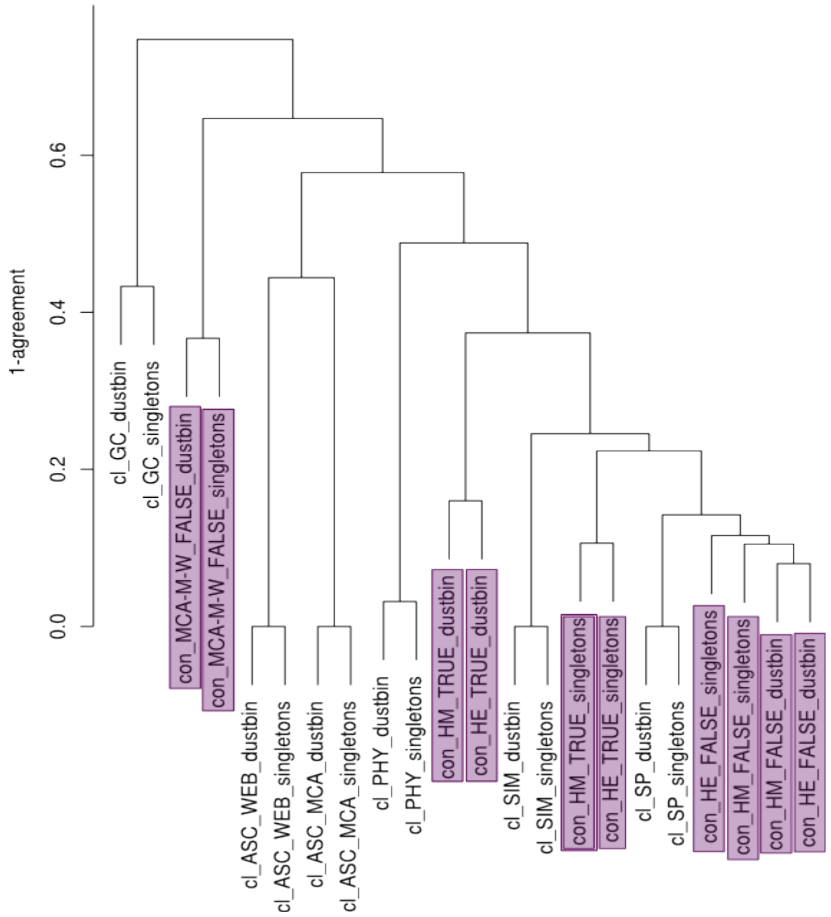


Illustration VIII.8: Agreement clustering tree

The agreement tree is a hierarchical clustering tree based on the disagreements between component and consensus clusterings. Component clusterings have the “cl_” prefix, and are: GC (Genomic Context), ASC_WEB (ASMC with CobWeb), ASC_MCA (ASMC with factorial analysis), PHY (PHYLOgenetic), SIM (similarity-based) and SP (Sci-Phy), each for different protocols for handling of unclustered proteins (dustbin or singletons). Possible consensus clusterings have the “con_” prefix and are boxed in purple. The different consensus clustering settings are: MCA-M-W (factorial analysis), HM (hard manhattan), HE (hard euclidean), for weighted and unweighted components (TRUE/FALSE, respectively), and for both unclustered proteins protocols (dustbin/singletons). The chosen MegaClustering corresponds to “con_HM_TRUE_singletons”.

Using this tree and details on the various consensus clusterings, the final parameter choices were:

Parameter	Choice	Justifications
SIM weight	0.66	All three of these methods were expected to share a large part of common information. Their weights were downsized to account for this and their respective coverages.
PHYLO weight	0.33	
SCI-PHY weight	0.33	
GC weight	2.00	The GC weight was favoured as it was built from a unique information source, and to compensate for its low coverage.
ASMC (CobWeb) weight	0.67	ASMC clustering is considered to be built on functionally-relevant amino acid residues, hence a relatively good weight, both clusterings considered.
ASMC (MCA) weight	0.50	
Cluster similarity measure	Manhattan	Manhattan distances are more stringent for integrated clustering construction than Euclidean distances, as they amplify differences.
Integrated clustering type	Hard	We wanted proteins to be each assigned to a single determined clusters, not probabilistically to several.
Singleton handling	Singletons	Empirically, the Dustbin approach tended to exaggerate false relationships between BKACEs whose only common point was that they weren't clustered by the same methods.

The chosen integrated clustering, affectionately called the **MegaClustering**, covered all 725 kept BKACE proteins and was composed of 32 MegaClusters, of which no singletons. MegaCluster size varied from 3 to 130, with an average of 22.7 and a median of 10.5. Only 4 MegaClusters contained more than 40 proteins (sizes: 56, 73, 99, 130), though 14 contained less than 10. This choice is comforted by a visual inspection of the agreement matrix below:

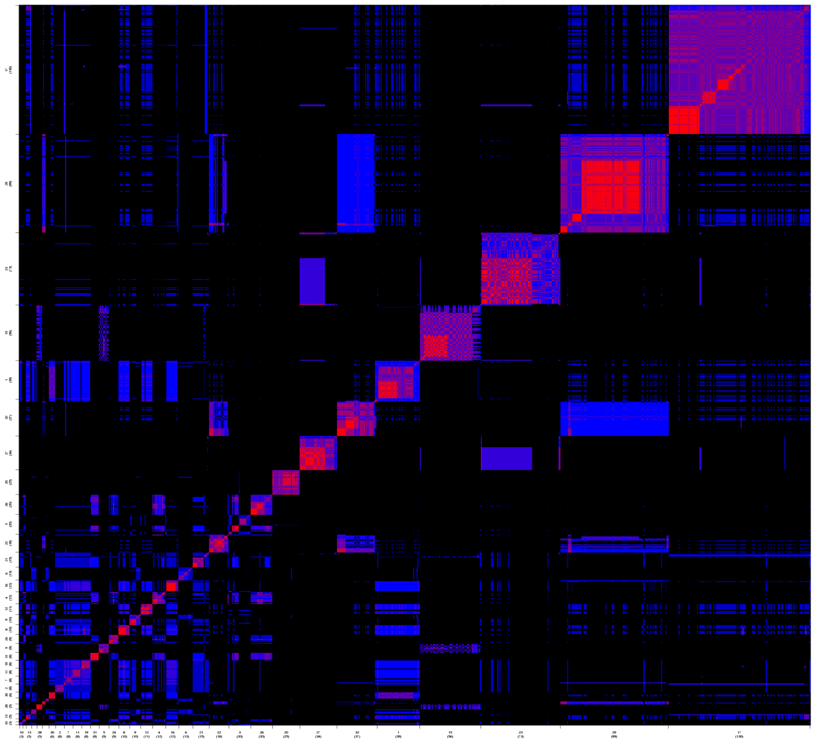


Illustration VIII.9: MegaClustering Agreement Matrix

The agreement matrix is a graphical representation of how well the different component clusterings agree, and how well this agreement is captured by the MegaClustering. For each pair (i,j) of clustered objects (here, BKACE proteins), the value in the matrix $M_{i,j}$ counts the number of distinct component clusterings that cluster objects i and j together. A colour scale from black to blue to red reflects in this figure the different value levels. Ordering the rows and columns according to increasing size MegaClusters should, in the case of high agreement between component and Mega clusterings, reveal a diagonal block-like structure of higher values, as is the case here.

VIII.C.7. Genomic Context Gene Families

In order to assist bioanalysis of individual BKACE proteins, it was decided that proteins from BKACE gene neighbourhoods should be clustered into gene/protein families that would capture information at the MegaCluster level. This was done using stringent BLAST score cut-offs to render binary all protein-protein similarities for BKACE context proteins. A similarity graph was then constructed using all BKACE *context* proteins for BKACEs from a same MegaCluster. A simple single-linkage algorithm (*i.e.* connected component detection algorithm) was used to isolate the families, which were then saved to the database. This very conservative approach should be useful in identifying only the best co-conserved gene families with the BKACE family, and should be instrumental in extracting specific annotations for them.

A total of 5,544 genes coming from all 32 GC clusters were clustered by this protocol into 955 gene families.

VIII.C.8. Inference of possible BKACE substrates

In [240], a subset of BKACE proteins was shown to catalyse the transformation of 3-keto-5-amino-hexanoate, and in [273] the study of the 3D structure of one of these proteins led to the proposition of a chemical mechanism that could be linked to key amino acids heavily conserved across DUF849. The working hypothesis behind the BKACE project was that the members of family DUF849 all catalysed similar reactions, alike in their chemical mechanisms, but differing in their substrates. More specifically, DUF849 proteins were expected to catalyse the cleavage of a carbon-carbon bond of 3-ketoacids (beta-keto acids, hence the name BKACE for beta-keto acid cleaving enzyme) for a variety of substrates with varying “chemical decorations” (*i.e.* chemical functional groups with different physical, chemical and steric properties) that other, less conserved amino-acids would interact with so as to define substrate specificity.

It was thus necessary to establish a list of potential substrates that could be catalysed by BKACEs. The most obvious approach would be to list all beta-keto acids known to life chemistry today. Using chemoinformatics tools proposed by KEGG and PubChem, we performed **substructure searches** for beta-keto acid groups, which lead to a long list that had to be manually pruned by our biochemists.

Not all these molecules were commercially available. Some were successfully synthesised by the Laboratoire de chimie organique et biocatalyse (LCOB) at Genoscope, though others were not. The list of substrates tested, along with their classification according to stereochemistry properties involved in substrate-enzyme interaction (as determined by some preliminary substrate docking experiments), and some indications of protocol modifications (for future reference[22]) are given below (chemical names have been simplified, and designate the substrate used in the biochemical assay, which does not necessarily correspond to the actual BKACE activity substrate due to protocol variations):

Name	Functional group	Charge / hydrophobicity	Steric hindrance	Protocol
Negative charged groups				
ketoglutarate	carboxyl	negative	+	
malonyl-CoA	carboxyl	negative	+	reverse
ketoadipate	carboxyl	negative	++	
succinyl-CoA	carboxyl	negative	++	reverse
Partial charged groups				
acetoacetyl-CoA	carbonyl	partial negative	+	reverse
hydroxybutyryl-CoA	hydroxyl	partial negative	+	reverse
Hydrophobe				
benzoyl-CoA	aromatic	hydrophobe	++	reverse
crotonyl-CoA	unsaturated aliphatic chain	hydrophobe	0	reverse
propionyl-CoA	aliphatic chain	hydrophobe	0	reverse
butyryl-CoA	aliphatic chain	hydrophobe	+	reverse
isobutyryl-CoA	branched aliphatic chain	hydrophobe	+	reverse
hexanoyl-CoA	aliphatic chain	hydrophobe	+++	reverse
decanoyl-CoA	aliphatic chain	hydrophobe	++++	reverse
methylketopentanoate	aliphatic chain	hydrophobe	+	ester
methylketohexanoate	aliphatic chain	hydrophobe	++	ester
methylketoisocaproate	branched aliphatic chain	hydrophobe	++	ester
methyloxooctenoate	unsaturated aliphatic chain	hydrophobe	+++	ester
Positive charged groups				
ketoaminohexanoate	aliphatic + amine	positive	+	
carnitine	aliphatic + amine	positive	0	carnitine
“Mixed”				
aminooxohexanedioate	carboxyl+amine	overall: neutral	++	

[22] Protocol modifications: “reverse”: reaction tested in reverse direction, *i.e.* from the CoA-ligated product; “ester”: substrate available as an ester, requiring coupling with an esterase reaction to produce, reaction tested in the normal direction; “carnitine”: this substrate required a specific protocol, reaction tested in the normal direction.

I shall now detail somewhat the experimental protocol used for testing these BKACE activities for these substrates.

VIII.C.9. Experimental Protocol

The details of the biochemical protocol(s) used to evaluate the activity levels of the BKACE proteins on the various proposed (and available) substrates will certainly be described in the BKACE article and are not part of my expertise. However, I shall underline several important points concerning the assays that determine in great part how I chose to conduct the statistical analyses of the results.

- **Choice of BKACE proteins to test**: many experimental factors influence the organisation and success rates of biochemical assays. The coding genes for BKACE proteins are generally GC-rich (*i.e.*, have a high proportion of Guanine and Cytosine in their nucleic sequence) which in practice means the genes are much harder to amplify by PCR before transferring them to expression vectors (transformed *Escherischia coli*). Furthermore, some of the BKACE host organisms were not available in the Genoscope prokaryote strain bank, nor commercially. BKACE proteins to test were thus manually selected by the LCAB team while trying to cover each MegaCluster evenly (roughly 10% of each cluster). 171 were finally successfully cloned and expressed in vectors.

- **Experimental design**: all chosen BKACE proteins (171) were tested against all available substrates (a total of 16 were tested). Crude total protein lysate was recovered from each vector expressing a different BKACE. The lysate was equally distributed across 4 wells for each substrate: 2 with the substrate, two without, making 2 repetitions for each. Furthermore, the entire experimental protocol was duplicated (the repetitions will be referred to as “repetition 1” and “ repetition 2”).

- **Activity estimation**: Activity was estimated using substrate-type-specific protocols across the 4 wells for each BKACE. These protocols had in common that they involved measuring the optical density (absorbance) at a given wavelength of each well over time, giving access to the concentration of a substrate or product during the experiment. The resulting curves were windowed and used to derive the initial reaction speed (which corresponds to the derivation of the curve at time 0, but in the case of BKACE activities, could be approximated by the slope of a linear function fitted to the windowed concentration curve).

Finally, the difference in initial reaction speeds between substrate-containing and substrate-free wells was taken to represent the initial reaction speed for the tested BKACE enzyme.

Due to the use of total protein lysate, in which the proportion of protein corresponding to the BKACE protein is unknown (and, as has been shown by the biochemists, is NOT constant across BKACEs), the actual activity measures are NOT directly comparable between BKACE proteins. This issue will be discussed in the statistical analysis sections.

VIII.C.10. Statistical representations and analyses

VIII.C.10.a. Repeatability analysis

The first step in validating the experiment requires showing that it is repeatable and not too subject to noise or exterior factors. To do this, I verified that measures were comparable from one repetition to the next. First, the repeatability of lysate quantities across repetitions is graphically represented in the figure below. The dots cluster around the diagonal, indicating good repeatability.

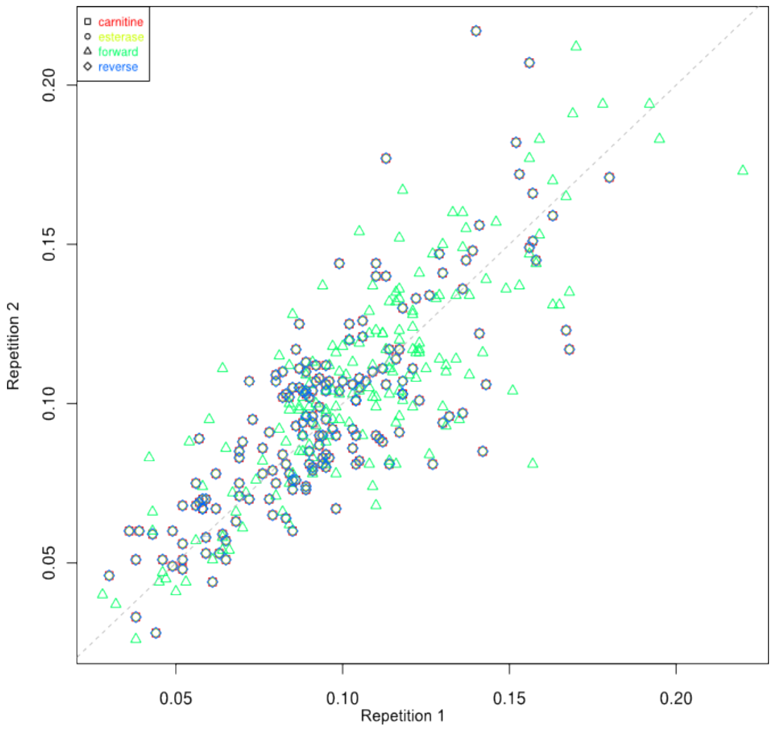


Illustration VIII.10: Lysate quantity repeatability

The quantity of total protein from extracted lysate is shown here, plotting values from repetition 1 against those of repetition 2 (arbitrary units). Even though it is not possible to link protein quantity to BKACE enzyme quantity, the relatively good repeatability of extracted protein is reassuring, allowing us to work with the hypothesis of comparable quantities of enzyme between repetitions. Different dot shapes correspond to different activity-measuring protocols.

Next, one can visualise enzymatic activity repeatability by plotting activity measures for one (substrate, BKACE) pair from one repetition against measures from the other. Again, the dots follow the diagonal, thus showing high repeatability.

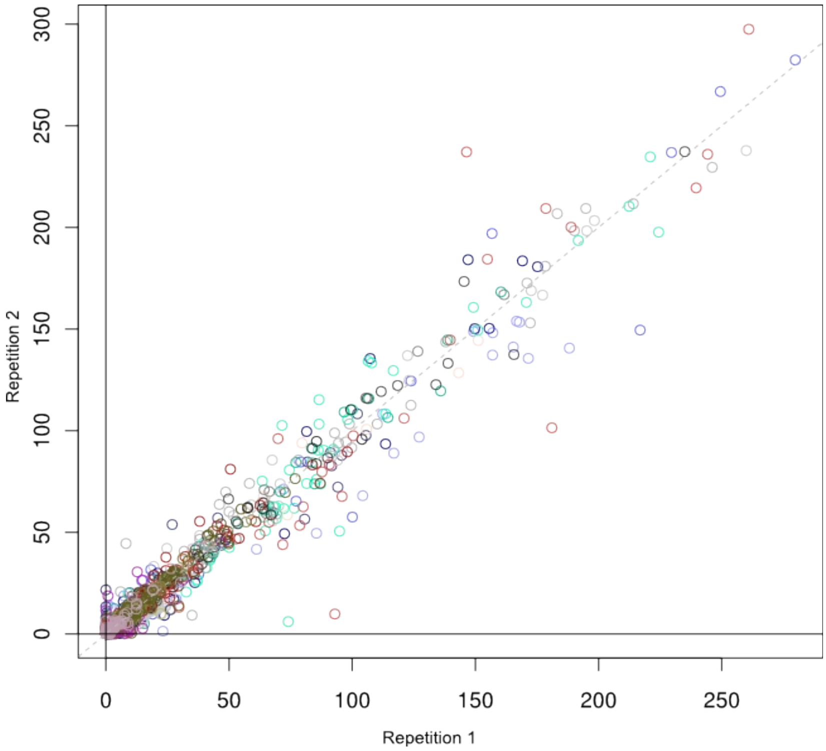


Illustration VIII.11: Activity repeatability

The measured activities are shown here, plotting values from repetition 1 against those of repetition 2 (arbitrary units). Each colour corresponds to one type of substrate. Negative activities are not shown. All points are heavily grouped around the diagonal, indicating that the activity measures are highly repeatable.

Taken together, these results reassured us that the measured biochemical activities were consistent across repetitions, allowing us to treat both equally. The next step in the analysis of these values is, of course, deciding which pre-treatments may be necessary.

VIII.C.10.b. Activity preprocessing

Biochemical assays are subject to many kinds of perturbations that add stochasticity to all measures. It is thus important to evaluate which measure values actually reflect the process measured, and which ones correspond to noise. This is particularly important in the BKACE project because, as said, different activity measures are not comparable between BKACE proteins. Some way, however, had to be found to distinguish active from inactive proteins for each substrate. I imagined and implemented several strategies for dealing with this particular kind of noisy data. I shall briefly describe different points of interest that are common or not to these different strategies, focusing more on the final choices made in the analysis for publication.

Manual cut-offs: The first, simplest idea is to ask the biochemists which activity measure values they consider as high enough to be trustworthy, and which ones they consider too small to be distinguishable from noise. This boils down to establishing a list of manually-chosen value cut-offs specific to each substrate. My first analyses were based on manual cut-offs supplied by the LCAB team.

Distribution-based cut-offs: The next idea was to analyse the activity measure distribution for each substrate, and attempt to base a cut-off on it. Manual definition of a cut-off on the basis of a distribution aimed at separating the numerous low values from the few high values, as only a minority of BKACE proteins were expected to act on any given substrate. I refined this procedure by fitting a mixture of normal distributions to the activity distribution using the “mixdist” R package. The idea behind this is simple: the measures we observe should arise from two parent distributions, one concerning the majority of proteins inactive on the given substrate and thus being centred on 0 or some small value, the other corresponding to the minority of active proteins for that substrate, but whose values cannot be compared, meaning that their parent distribution is a very wide normal distribution[23] centred on the average measure value. Proteins were then declared active or not when their activity measure was over the value for which the inactive distribution becomes negligible (empirically 10x less) when compared to the active one. The final approach kept for publication was this distribution-based cut-off procedure.

Real values *versus* Binary values: Once a cut-off has been applied, the activity measures can either be kept as is (*i.e.* zero or strictly positive real values) or rendered binary (0 or 1). Most of my first analyses were based on real-valued activities, though since values were not comparable between BKACEs, we finally decided to use binary values as whatever resolution was lost in this transformation could not correspond to any actual difference in activity levels, but only in BKACE protein expression levels.

Normalisation: Under the hypothesis that the fraction of protein in the total lysate corresponding to BKACE proteins was constant across BKACEs, “normalising” (dividing) measured activities by the total quantity of complete lysate was an obvious procedure for results to be comparable.

[23] It would have been interesting to try out a uniform distribution rather than an extremely flat normal distribution. However, estimating the borders of a uniform distribution would not have been straightforward, and I was unable to find an R package capable of estimating a mixture of distributions of different types (normal + uniform).

However, given that the fraction was NOT constant across BKACEs, it remained debatable whether to normalise or not. I thus tested with and without. Finally, we decided that it was not useful to normalise as it had very little influence in practice on numerical or binary activities.

Profiling: Activity measures might not have been comparable between BKACE proteins, but they were across all substrates for a given BKACE. Hence the idea of analysing not the activity values themselves, but the fractions of each observed activity divided by the total activity per BKACE. Activity measures, expressed as percentages, thus become comparable across BKACEs. Unfortunately, this approach has several drawbacks:

- percentage analysis is not as straightforward as value analysis, especially for multivariate statistical methods such as ANOVAs and Factorial Analysis;
- activity percentages behave in a particular way in respect to total activity. Thus, for proteins with systematically low activity values (*i.e.* not responsive to any substrates), the small values that subsist (even after applying the cut-offs if any) and that correspond to noise are over-represented;
- working with profiles based on binary values only worsens the previous problem;
- a profile corresponding to no values (when none passed the cut-off) is meaningless.

I experimented heavily with profiling as it was the only way to render measures comparable across BKACEs. Finally, however, working with simple binary activities was simpler, more intuitive and easier to analyse correctly than working with profiles.

Perspectives: I am relatively sure that a better approach based on distribution mixtures and fuzzy clustering is possible to process this data with better resolution and power. It is too late, however, to try to develop a new analytical method for the upcoming publication. Once the next similar project is initiated, I intend to propose a collaboration between the Genoscope and the Laboratoire Statistiques et Génome, where I did my masters internship, whose expertise includes dealing with this kind of data.

VIII.C.10.c. Substrate tree and clustering

One might wish to compare the pre-defined substrate typology to the observed results, in the hope of determining if there is any link to activity profiles and substrate characteristics. It is possible to build a hierarchical tree based on the similarities between their activities as measured across all tested BKACE proteins. To build such a tree, I carried out a ascending hierarchical clustering using "Manhattan" distances (in order to exaggerate differences) between binary substrate activity profiles, though other protocols could be used. This tree can also be cut in order to establish a clustering of substrates that might then be interesting to compare to the substrate typology based on physico-chemical properties. Unfortunately, I did not have the time to deepen this king of analysis, as it would have required an improved formalisation of the substrate typologies that I was not qualified to make. The tree itself, however, was useful for the graphical representations, as described below.

VIII.C.10.d. Graphical representation of results

In order to appreciate how the BKACE proteins performed across substrates, hopefully showing that the MegaClustering (or any component clustering of interest) indeed segregates proteins into iso-functional sub-families, it was necessary to conceive graphical representations that could account for the different factors. Several points relative to the representation are discussed here.

Substrate typology: substrate typology was determined *a priori* on the basis of substrate physico-chemical properties, as detailed in section VIII.C.8. Substrate typology can be represented by a colouring scheme. Activity-based substrate clustering could also be representative of substrate typology, and can be represented by a hierarchical tree of substrates, as described in the previous section.

BKACE phylogeny: BKACEs can be organised according to any evolutive phylogeny. Here we have access to the QuickTree-based phylogeny established on the basis of BKACE sequence similarity. We also had access to the ASMC tree that is a sort of phylogeny based on the 3D positions of key active site amino acids, though no figures using this tree will be shown here.

Component/Mega Clustering: BKACEs can be organised or coloured according to component/consensus cluster belonging.

Real/Binary values: different real activity values for a same BKACE can be represented by stacked coloured bars. Binary activities can be represented by stacked bars or by a presence/absence matrix.

These different considerations led to various representations, of which the most informative and visually striking are detailed hereafter.

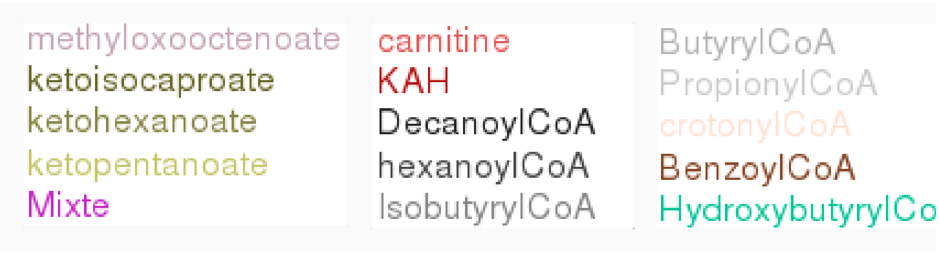


Illustration VIII.12: Substrate colour code chosen for BKACE graphical representations

Each colour corresponds to a single substrate. “Mixte” designates aminooxohexanedioate, and KAH designates ketaoaminohexanote, which is the original substrate from the lysine degradation pathway. This colour code will be used in all figures for the substrates.

Real-valued activities (with cut-off) grouped by MegaCluster:

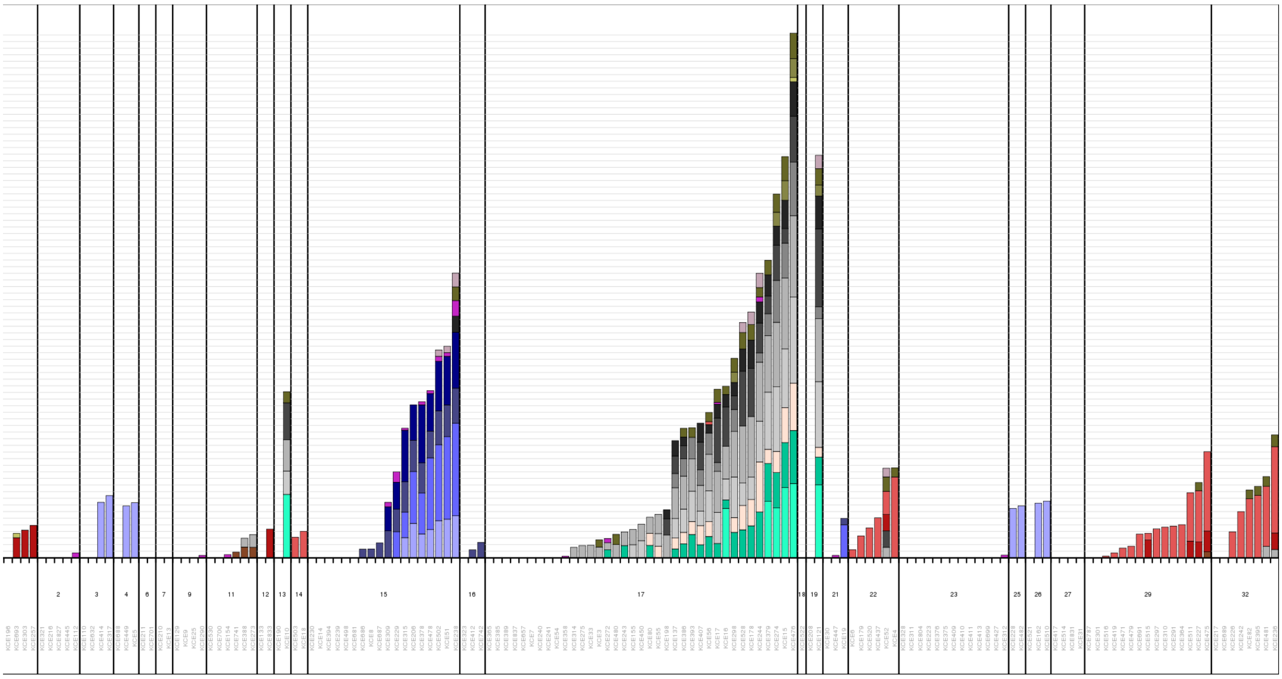


Illustration VIII.13: Real-valued activities (with cut-off) grouped by MegaCluster

Each colour corresponds to a single activity. Each bar corresponds to the stacked activities of each BKACE. BKACEs are horizontally grouped according to their MegaClusters. As activity measures are not comparable between BKACEs, the vertical axis can be considered to have an arbitrary unit.

Real-valued activities (after applying distribution-based cut-offs) for each BKACE are represented as stacked bars coloured according to substrate typology. BKACE are horizontally organised according to MegaCluster membership. It is visually obvious that generally, each MegaCluster has a coherent activity profile. This is a first step in validating the MegaClustering approach, as well as the working hypothesis: BKACEs can be organised into sub-families that catalyse mechanistically-similar reactions on different substrates.

Binary activities organised by phylogenetic and substrate trees:

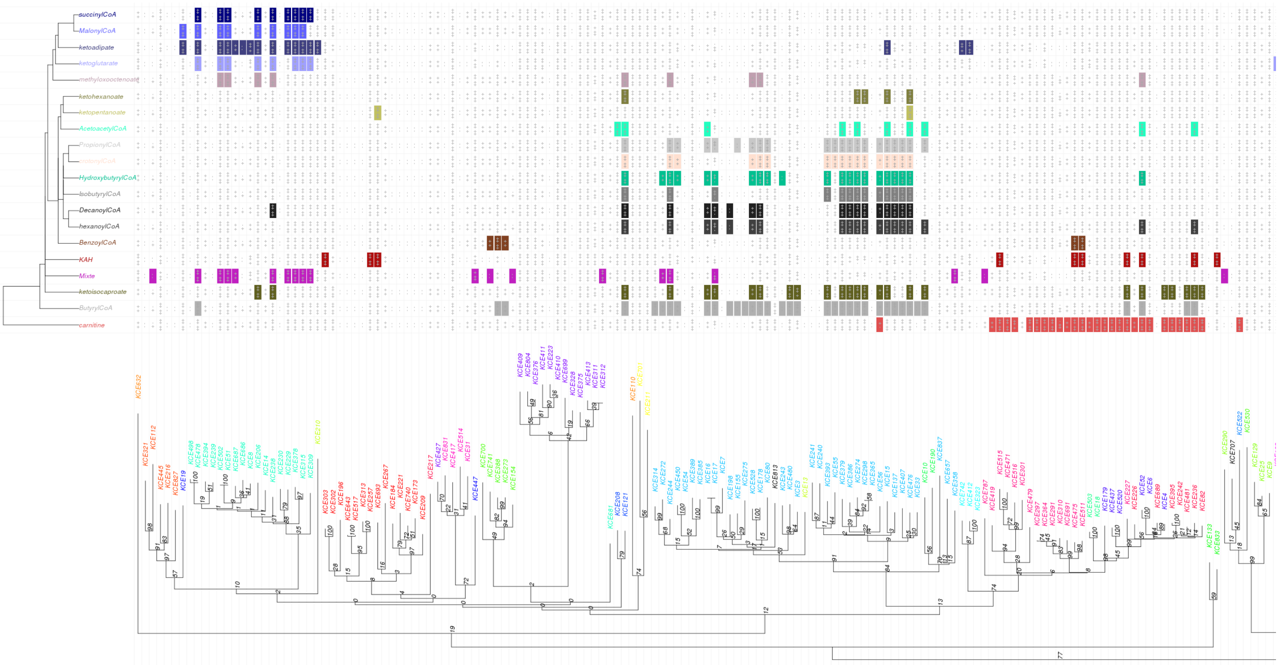


Illustration VIII.14: Binary activities organised by phylogenetic and substrate trees

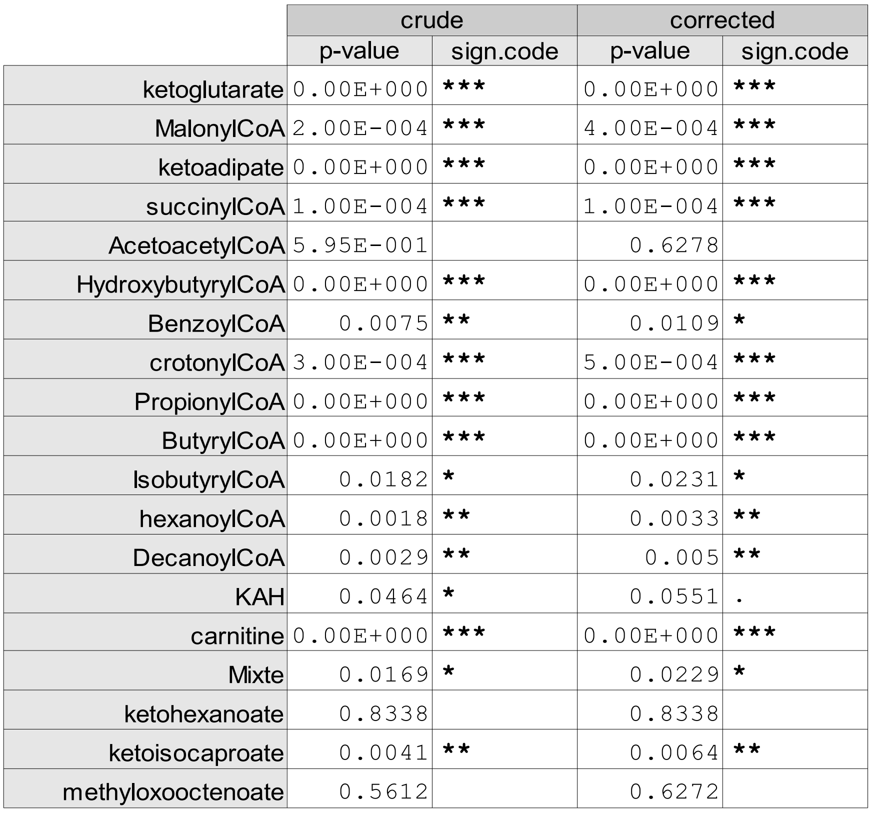
Each row in the matrix corresponds to a substrate, and each column to a BKACE protein. Non-white blocks indicate a positive activity. Substrates are organised according to a hierarchical tree built on binary activity similarities. BKACEs are organised according to the corresponding phylogenetic tree below. The leaves of the phylogenetic tree are coloured according to MegaClustering membership. "+" and "-" signs are qualitative indications of the quantity of protein extracted in the experiments: "-" values are to be interpreted with more care (as in these cases, the absence of an activity could correspond to the absence of the protein, rather than the absence of catalysis).

Binary activities (after applying distribution-based cut-offs) are represented by squares coloured according to substrate typology. Substrates are organised vertically according to the hierarchical tree based on activity profile similarity described in section VIII.C.10.c. BKACEs are organised horizontally according to phylogeny. The phylogenetic tree leaves are coloured according to MegaClustering. Once again, this graphical representation illustrates the coherence between an *a priori* grouping of BKACEs (here, the phylogenetic tree) and the observed activity profiles.

VIII.C.10.e. Statistical validation of clustering

Validating the MegaClustering in a rigorous, statistical manner, in respect to the biochemical results, is at best, a hard task. Indeed, one must compare a partition of the proteins in the DUF849 family, to the protein's respective activities, or some sort of clustering thereof. Several protocols could be imagined, each with their advantages and drawbacks.

The first idea I put into application after several discussions with Alain VIARI was to compare, for each separate activity, the partitioning of BKACE proteins according to a) the MegaClustering and b) the presence/absence of said activity in the proteins. Fisher's exact test is adapted for this kind of comparison. Indeed, it tests whether the modalities of two categorical variables (here, the different MegaClusters, and the presence/absence of the target activity) “attract” or “repel” each other, *i.e.* some combinations of each are seen more (respectively less) often then expected by chance alone. I thus conducted a Fisher test on each activity, and used the Benjamini-Hochberg correction for multiple tests. The results are summarised in the table on the next page.



	crude		corrected	
	p-value	sign.code	p-value	sign.code
ketoglutarate	0.00E+000	***	0.00E+000	***
MalonylCoA	2.00E-004	***	4.00E-004	***
ketoadipate	0.00E+000	***	0.00E+000	***
succinylCoA	1.00E-004	***	1.00E-004	***
AcetoacetylCoA	5.95E-001		0.6278	
HydroxybutyrylCoA	0.00E+000	***	0.00E+000	***
BenzoylCoA	0.0075	**	0.0109	*
crotonylCoA	3.00E-004	***	5.00E-004	***
PropionylCoA	0.00E+000	***	0.00E+000	***
ButyrylCoA	0.00E+000	***	0.00E+000	***
IsobutyrylCoA	0.0182	*	0.0231	*
hexanoylCoA	0.0018	**	0.0033	**
DecanoylCoA	0.0029	**	0.005	**
KAH	0.0464	*	0.0551	.
carnitine	0.00E+000	***	0.00E+000	***
Mixte	0.0169	*	0.0229	*
ketohexanoate	0.8338		0.8338	
ketoisocaproate	0.0041	**	0.0064	**
methyloxooctenoate	0.5612		0.6272	

Illustration VIII.15: Fisher tests on binary activities versus MegaClustering

For each substrate, results of the Fisher test are given, pre-correction (“crude”) and post-correction using Benjamini-Hochberg. Significance codes ('***': highly significant, '**': moderately significant, '*': significant, '.': non significant but could be a trend) are shown for convenience.

Even after correcting for multiple tests, most substrates have their binary activities either strongly associated or strongly dissociated with MegaClusters. This suggests that the MegaClustering does indeed capture some measure of substrate specificity amongst the BAKCE proteins.

My second idea involved using factorial analysis to try and bridge the gap between multiple variables (activities) and one categorical variable (the MegaClustering partition). A Principal Components Analysis on the real-valued activities (or a Multiple Correspondence Analysis on the binary values) could distribute the proteins into a multidimensional space, within which one could either 1) establish an activity-based clustering of proteins, to be compared to the MegaClustering, or 2) project the MegaClustering as an illustrative variable, and evaluate visually if proteins of similar activities correspond to the MegaClusters. The latter is of little more use than the previous validations, as it requires manual expertise of the results as well. I chose to cluster BKACE proteins using their binary activity profiles, using an MCA and a hierarchical clustering on a Manhattan distance in the factorial space. After manual inspection, the obtained tree could be cut cleanly into 9 clusters. I thus attempted to apply Fisher's test to this clustering and the MegaClustering. Unfortunately, this proved to be non-tractable, and I resorted to an approximate Chi-Square test. The test concluded that there were highly significant attractions/repulsions between activity-based clusters and MegaClusters (p-value < $2.2*10^{-16}$). Manual inspection of the contingency table shows this as well, though the segregation is not as exclusive as we might have hoped (*i.e.* each MegaCluster corresponding to one activity cluster and vice-versa). To illustrate the cross-links, I generated a graphical representation of the contingency table, shown below:

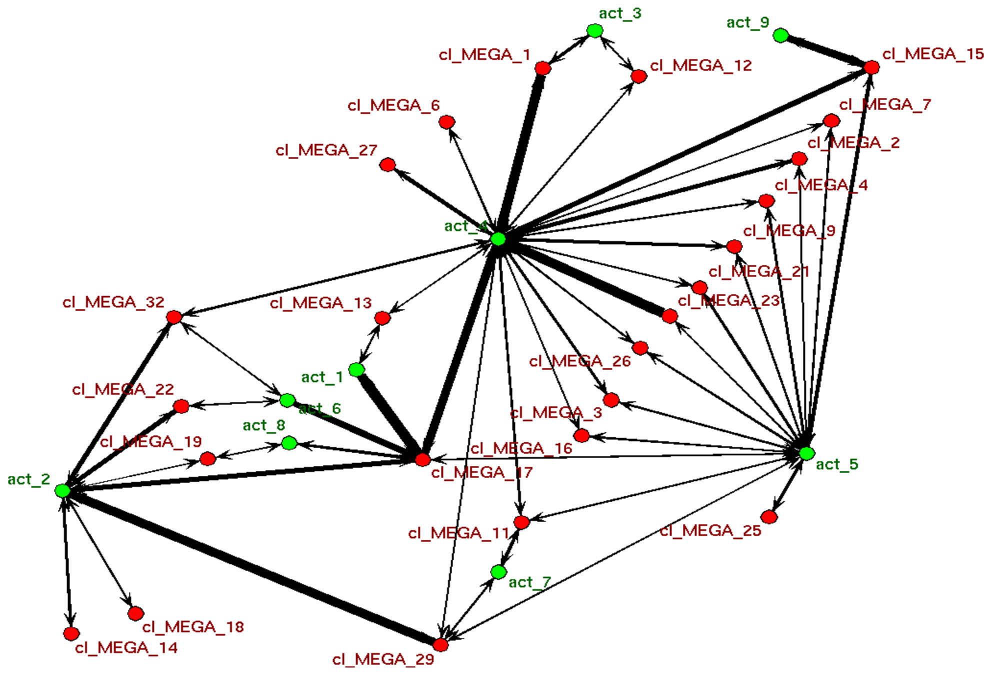


Illustration VIII.16: Correspondence between activity-based clustering and MegaClustering

The MegaClustering contained 32 clusters (cl_MEGA_1 to cl_MEGA_32), the binary activity-based clustering had 9 (act_1 to act_9). Correspondences between the two are represented here in a graph, whose adjacency matrix corresponds to the contingency table between the two clusterings. This means that the thickness of an edge between a MegaCluster node and an activity cluster node is proportional to the number of BKACE proteins that is contained by both. Despite multiple correspondences between clusters of each type, the BKACE proteins are not distributed as one would expect by pure chance.

Another idea that I chose not to explore would have been to use Linear Discriminant Analysis to see if it would be possible to discriminate MegaClusters on the basis of real-valued activities. An issue with LDA is that it attempts to build linear predictors, based on the quantitative variables (activity measures), that can be used to maximally separate cluster barycentres. As already said, the real-valued activities are not really exploitable due to the impossibility of comparing numeric results across BKACEs. Furthermore, such linear predictors would not correspond to latent biological concepts, and their interpretation would thus be near impossible. For these reasons, I did not carry out this particular analysis.

VIII.D. Conclusion

In this study, a collection of bioinformatics strategies and experimental procedures allowed the exploration of the functional space of a yet little-known enzyme family called DUF849. A comparative genomics study suggested a first enzymatic activity for a dozen DUF849 proteins: the 3-keto-5-aminohexanoate cleavage enzymatic activity (KCE) from the lysine degradation pathway, which was confirmed by biochemical assays. The experimental determination of the 3D structure of the protein opened the way to establishing the chemical mechanism of the KCE activity, and of its extrapolation to the generic BKACE activity for the rest of the family. In order to explore this avenue, protein members of DUF849 were clustered together using both sequence- and context-based methods, in order to obtain sub-families that would hopefully be iso-functional in respect to substrate specificities. Potential candidate substrates for the BKACE activity were selected using chemical substructure searches. The sub-families were used to select BKACE proteins for biochemical testing. Activity profiles seemed to concord well with sub-family delimitation. All in all, the close collaboration between bioinformatics and experimental biochemistry has helped advance our understanding of DUF849, and has led to the discovery of several novel enzymatic activities of type BKACE. This strategy might be applicable (warranting some modifications) to the study of other families of unknown function, or even of families of known function but suspected of functional promiscuity and/or underground metabolism.

I shall now discuss some of our results, in the light of work still in progress.

VIII.E. Discussion and perspectives

VIII.E.1. From *in vitro* to new metabolic pathways

The work on the BKACE project has led to the discovery of many novel enzymatic activities. To obtain additional information about these, we conducted a **bioanalytical survey of the genomic contexts** of BKACE proteins with confirmed activities. The rationale behind this was, if bioanalysis traditionally found functional clues from the functions of co-localised genes, then conserved co-localised genes should be all the more informative. Three genomic and metabolic contexts were validated by manual bioanalysis: the confirmation of the original lysine degradation pathway (which is already presented in [240]), and the discovery of two new pathways concerning beta-ketoadipate and dehydrocarnitine cleavage activities.

Ketoadipate: The ketoadipate degradation pathway in MetaCyc (ID: PWY-2361) is composed of two reactions. The first transfers a CoA to the ketoadipate. The second uses another CoA while breaking a carbon bond of the ketoadipyl-CoA, producing succinyl-CoA and acetyl-CoA. From a carbon chain point of view, the ketaoadipate BKACE activity summarises both these reactions in a single step, while using one less CoA.

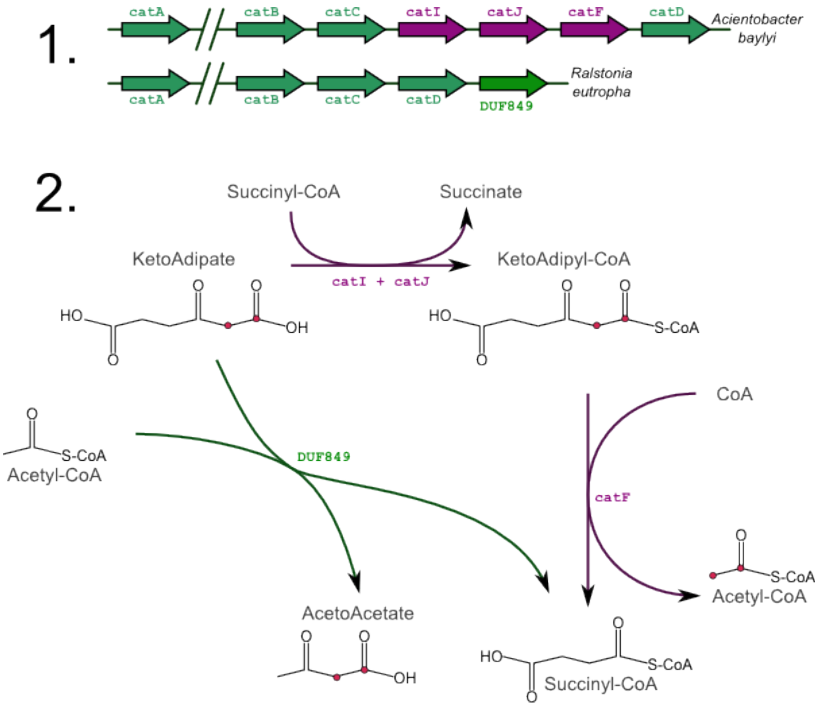


Illustration VIII.17: The BKACE ketoadipate degradation pathway

1. Ketoadipate degradation operons in *Acinetobacter baylyi* and *Ralstonia eutropha*.

2. Ketoadipate degradation pathways. In *Acientobacter baylyi*, ketoadipate is transformed into succinyl-CoA and acetyl-CoA via two reactions catalysed by genes catI, catJ, and catF. In *Ralstonia eutropha*, a single DUF849 protein transforms ketoadipate into acetoacetate and succinyl-CoA.

Dehydrocarnitine: In MetaCyc, the carnitine degradation pathway variant II (ID: PWY-3602) involves the transformation of carnitine into dehydrocarnitine by a dehydrogenase, before being processed by two hypothetical decarboxylation reactions into glycine betaine. With a DUF849 enzyme, these two reactions could be replaced by two others: the BKACE activity, and a thioesterase to transform the produced glycine betaine-CoA into glycine betaine. In *Pseudomonas putida*, a potential operon (genes with MicroScope labels PP0301 to PP0303) containing a carnitine dehydrogenase, a DUF849 member and a putative thioesterase was found. The biochemical assays later showed that the DUF849 member indeed catalysed the cleavage of dehydrocarnitine. The activities of all the gene products for this pathway have now been experimentally confirmed (data not shown).

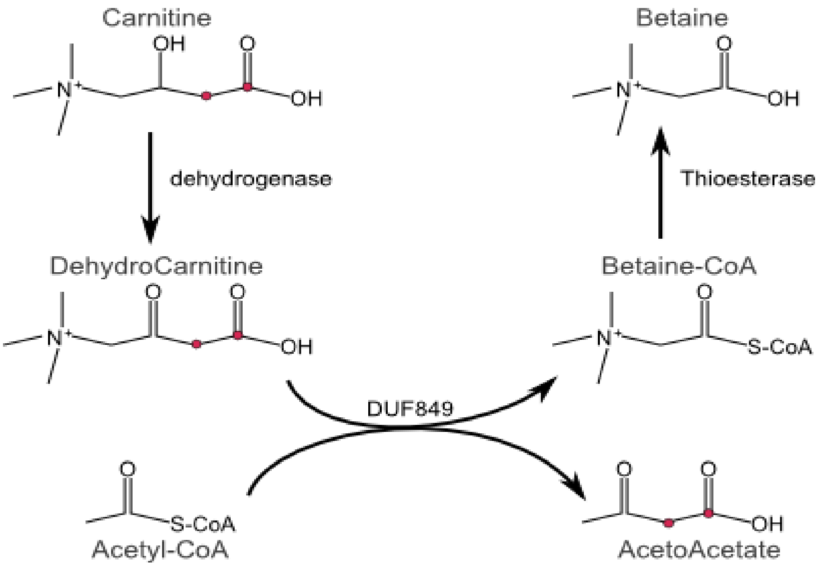


Illustration VIII.18: The BKACE carnitine degradation pathway

Carnitine is transformed into dehydrocarnitine (a beta-keto acid) upon which the DUF849 protein acts to generate betaine-CoA and acetoacetate. The former is then processed by a thioesterase into betaine. The three genes encoding the enzymes from this carnitine degradation pathway were found in a single *Pseudomonas putida* operon.

Thus, our study led us to the discovery of novel enzymatic activities, of which some were assignable to new metabolic pathways. Associating these pathways with operons proves that the pathways are of biological relevance to the host organisms.

VIII.E.2. High-throughput Screening limitations

The high-throughput screening protocol briefly described in this manuscript allows the simultaneous experimental validation of the presence or absence of many activities across a large number of proteins. It does, however, have several drawbacks, some of which have been pointed out previously: low comparability of results, imprecision in substrate specificity by having no access to traditional enzymology values (K_M, V_{max}... which could better describe the enzymatic efficiency), necessity of using an expression vector that can bias the results... Any biochemist wishing to publish enzymology data would thus have to carry out separate characterisation experiments. This is not necessarily a problem, as the high-throughput screening was designed to be purely exploratory, essentially greatly reducing the amount of characterisation experiments to perform.

An alternative for the evaluation of substrate specificity within BKACE subfamilies would be the use of homology modelling and **3D docking simulations** to check which BKACE proteins can accept a given substrate into their active site. Karine BASTARD (from the previous LABIS team) has initiated such a study, and preliminary results have help deepen our understanding of the interaction between substrates and BKACE proteins. Once again, combining bioinformatics analysis with experimental tests has proven to be beneficial.

VIII.E.3. Genomic Context Clustering

As the definitions of genomic context and the methods for dealing with it are varied, the ways of generating clusterings based on them are numerous as well, though -to our knowledge- not so actively researched. It may thus be possible to come up with different bioinformatics protocols in order to generate genomic context clusterings.

Multi-genome syntons could be seen as a means of generating clusters of similar neighbourhoods [5]. Each mutli-genome synton would thus be a genomic context cluster, though some way of dealing with overlapping syntons would have to be found. Other, similar approaches to this are presented in [143,289]. In comparison, our approach is much simpler than these, requires less data, is faster to run and interpret, even though it requires more manual work. This is consistent with our very “hands-on” *modus operandi* in this study. And considering that the GC clustering is only a component of the MegaClustering, as well as a crude basis for finding genomic contexts for target BKACE reactions, it might be considered overkill to develop a more complicated method. This choice will be up to future users of BKACE-inspired strategies.

VIII.E.4. Why a BKACE activity ?

Biochemical reactions are always of interest to many industrial applications (in phamacology, cosmetics, agronomy, agrochemistry, food industry...) as they often offer many advantages in respect to pure chemistry approaches; a panorama of the place of biochemistry in industry can be found in [231].

The BKACE reaction may not appear, at first sight, of any particular industrial interest. Involved chemical entities are relatively common, though they do not participate in any metabolic pathways that draw any attention for the time being. The decarboxylation of beta-keto acids into shorter carboxylic acids is known to occur spontaneously in normal working lab conditions, so one may wonder at the use of an enzyme capable of generating these products. The use of a BKACE enzyme, particularly for its host, lies in the fact that rather than producing carbon dioxide and a carboxylate via the spontaneous reaction, it produces acetoacetate and a carboxylate activated by a Coenzyme-A group (see [Illustration VIII.1: The lysine fermentation pathway] page 157). Both the latter are much more easily re-injected into the organism's central metabolism. A hypothetical gain in fitness brought on by possessing a BKACE protein is highly supported in some organisms due to the existence of genomic contexts surrounding its coding gene, even replacing a two-step reaction by a single-step reaction in the case of ketoadipate. Likewise, BKACE might interest the industry, if this kind of metabolic pathway optimisation is required. However, the objective of this study was not to revolutionise the biochemical industry, but to prove the worth of the presented investigation strategy, which can now be applied to other families of higher technological value.

VIII.F. Conclusion

A comparative genomics breakthrough, confirmed by biochemical assays, allowed researchers at the Genoscope to gain a foothold in the functional space of metabolic reactions catalysed by proteins of the BKACE family. Analysis of the corresponding conserved domain of unknown function (DUF849) and of the genomic contexts of said proteins suggested that a family of similar metabolic reactions was waiting to be discovered. The strategy described here was created in order to explore this diversity of reactions through a close association between bioinformatics-based hypotheses and high-throughput experimental verification. Several of the molecules proposed by the strategy were effectively found to be substrates for BKACE proteins, validating the proposed mechanisms (though not all are perfectly understood). The high-throughput approach, despite several drawbacks, was able to confirm or refute these activities within the set of tested proteins, thus narrowing down future, more biochemically precise tests. Finally, in both manual and statistical tests, the sub-families generated by bioinformatics analysis appeared to be relatively homogeneous in terms of metabolic functions, thus validating the entire strategy.

Thanks to this study, a set of beta-keto acid cleaving enzymes corresponding to previously-unknown enzymatic activities has been elucidated. We have thus validated our strategy on a previously unknown family. It should be possible to adapt and improve upon this strategy in order to explore the functional diversity of other unknown families for which some functional starting point is available. Furthermore, it should be possible to rework our visions of *known* families with known functions, refining our understanding of the involved mechanisms, and describing with increased precision the functional promiscuity of the enzymes. Indeed, as has already been pointed out, our knowledge of functional promiscuity is expected to be far from complete, due to the loss of interest a protein has once one of its functions have been discovered. It is our hope that this strategy will be of use in extending knowledge of enzyme families far and wide.

IX. Other works

IX.A. The Carnoulès Acid Mine Drainage project

IX.A.1. Introduction

As presented previously in this thesis, “metagenomics” englobes techniques and analyses that allow the sequencing and study of multiple genomes at once. It is of particular interest when dealing with complex ecosystems containing microbial species for which pure cultures have not been obtained *in vitro*. This study focused on the genomes and metabolic capacities of the microbe ecosystem that has developed in a particular extreme environment: the acidic, sulphurous and arsenic-rich drainage waters of a mine in Carnoulès, known to pollute the stream Reigous (Gard, France). Previous studies underlined the role of the said ecosystem in the cleansing of Reigous waters over the 1.5 km run until its confluence with the river Amous, where sulphur and arsenic levels drop to 5% of their initial values. The metagenomic analysis of the Canoulès Acid Mine Drainage was designed to explore the genomes and metabolic capacities of the microbes responsible for this clearance. The article published on this project is available as an annex to this manuscript.

IX.A.2. Work provided

Damien MORNICO was assigned to working on the predicted reaction content on the Carnoulès supercontig bins. His objective was to compare the different bins in order to establish common features (with the idea that these might be linked to adaptations to living in the extreme Carnoulès conditions) and specific features (with the idea on establishing possible ecological relationships between different species), hopefully generating a revealing figure that would put forward these notions.

Metabolic reaction content of the various bins was predicted using a) the automated transfer procedure on the basis of sequence homology, and b) domain-based predictors, that are part of the main MicroScope pipeline. As we were dealing with a metagenome, for which the assembly was still as unverified bins, we decided to analyse metabolic reaction absence/presence across the different bins. This would give us a high-level view of the metabolic capacities of the various bins, their specificities and common points (and perhaps help confirm them as coming from separate organisms).

The high-dimensionality of the data, as well as the specific objectives of this analysis, suggested use of factorial analysis. I worked with Damien MORNICO on the analysis of his metabolic data. All figures and results are available in the paper itself.

IX.B. The Bradyrhizobium project

IX.B.1. Introduction

Bacteria of the *Bradyrhizobium* genus are common micro-organisms that live in soil and that are able to colonise legume roots and stems, forming symbiotic nodules. The project presented here focused on the comparative analysis of the genomes of 9 *Bradyrhizobium* strains specific to legumes of the *Aeschynomene* genus, selected for the atypical characteristics of its symbionts. In particular, some of these strains have photosynthetic capacities; some are able to colonise plant stems as well as roots; and some enter plant symbiosis without using the well-known *nod* gene-dependent pathway. The general objective of the study was to reveal how *Bradyrhizobia* have adapted to their ecological niche, and how the specificities of the selected strains arose. The article published on this project is available as an annex to this manuscript.

IX.B.2. Work provided

The work provided for the *Bradyrhizobium* project closely resembles that of the Carnoulès project; indeed, the same kind of metabolic analysis is performed in both articles. In this case, however, the studied bacterial strains were not sampled from a whole ecosystem, but are closely-related strains from a given phylum. Furthermore, additional information was available for the strains: their specific abilities in respect to their symbiosis with host plants. This analysis thus focused on associating metabolic pathways to the latter. Finally, the metabolic information analysed was different: as we were using high-quality rebuilt genomes, this study analysed metabolic pathway completion values rather than simple reaction absence/presence. The pathway completion of a pathway P in organism O is the number of reactions of P that are known to be catalysed in O over the total number of reactions in P. This measure can give an idea of the evolutive pressures applied to an organism's metabolism (*i.e.* high completion for important pathways, low completion for pathways lost or in the process of being lost). Covariance of this measure between pathways across phylogenetically related organisms can be indicative of common pressures, especially in a high-level analysis, though caution must be observed when interpreting fully completed pathways, or totally absent pathways. Another problem could be the existence of pathway variants amongst studied strains, which can break pathway completion covariance even though a same evolutive pressure exists. Finally, when observing factorial planes generated by analysing percentages (such as pathway completions), one must not forget that the dot cloud is confined to a hyper-cube or a hyper-pyramid. All figures and results are available in the paper itself.

X. Global discussion & perspectives

X.A. Overview

Four projects that I worked on during my thesis were presented in this manuscript. The CanOE strategy contains a method for locating genomic metabolons, units of metabolic function of prokaryote genomes, that are then exploited in a novel, multi-organism way in order to generate hypothetical gene annotations with scores that can help bioanalysts evaluate their worth, an approach that is particularly interesting when the involved metabolic reaction is a sequence orphan one. The BKACE project focused on exploring the functional space of a protein family of previously unknown function, using bioinformatics methods to generate iso-functional subfamilies and potential alternative reaction substrates. These hypotheses were then tested in a high-throughput enzymology screening, an original development in itself by the Genoscope's LGBM, LCOB, and LCAB teams. Finally, two different applications of factorial analysis to metabolic data concerning metagenomes or related genomes were used to explore the metabolic specificities and generalities of multiple prokaryote organisms at once.

X.B. Discussion

The developments presented in this thesis all deal with functional annotation - more specifically, metabolic annotation - of prokaryote genomes. They exploit contextual information, established by comparative genomics approaches, in order to propose metabolic activities for protein-encoding genes. These developments differ greatly in objectives, data and methods, and as such can be used in complementary ways.

X.B.1. The scope of each project

The Factorial Analysis carried out in Carnoulès project allowed the study of the metabolic relationships between micro-organisms in a community, while that of the Bradyrhizobia project shed light on the evolutionary specificities of micro-organisms from a same genus. These statistical analyses are, in essence, high-level explorations of the metabolic capacities of the provided genomes. They were designed not to discover novelty nor to generate annotations, but to provide a general picture of the workings of the studied organisms that is essential for any bioanalyst wishing to continue annotating his or her genomes.

The strategies implemented in the CanOE and BKACE projects were designed to generate annotations. CanOE can propose candidate genes for enzymatic activities in each prokaryote organism analysed, and can thus be seen as a "horizontal" annotation tool. It does, however, exploit gene family information, allowing it to propose candidate families for activities. The gene families are also instrumental in ranking candidate gene/family propositions. CanOE is thus also a "vertical" annotation tool, as it can cover multiple organisms at once. The BKACE project is also a "vertical" annotation tool, as it studies a single target gene family at a time. It does include a "horizontal" component, as sub-family-level annotations were proposed using integrated local genomic context and protein 3D structure studies.

In summary, each of the projects I worked on in this thesis generated different-level views useful to genome annotation: high-level for the metabolic factorial analyses, multi-organism gene families for BKACE, and cross-organism annotation for CanOE.

X.B.2. CanOE & BKACE are complementary

As underlined in the previous section, the CanOE and BKACE projects are complementary in their approaches to genome annotation. In practice, it is obvious that a single hypothetical annotation generated by CanOE between a candidate gene/family and a metabolic activity (be it global orphan, local orphan, or not as in the case of annotation policing) can serve as a starting point for a brand new BKACE-like analysis, focused on the biochemical declinations of the target activity. This will be especially important for families with evidence of being associated to several, similar reactions. Due to its pragmatism-driven design, the family-generating algorithm of the CanOE approach might not be the best adapted to delimiting a candidate gene family for examination: it might propose several small families at once, one big low-specificity family with many sub-families, or anything between these extremes. It would then be necessary to use the CanOE families as starting points for the identification of more accurate families, by searching public databases (such as Pfam) for correspondences. In the alternative CanOE multi-genome integration procedure presented in section VII.H.4.d, families are not used at all; only a set of candidate genes with varying confidence scores would be retrieved for a target metabolic reaction. Again, in would then be necessary to search existing family databases for families encompassing these candidates in order to launch a BKACE-like study. It would theoretically be possible to base a BKACE-like study on a group of proteins that do not belong *a priori* to a given family, though this would be risky and would make the entire procedure more complex and doubtful. Indeed, several steps of the BKACE strategy presented here rely on the construction of a Multiple Sequence Alignment of the involved proteins (family pre-filtering, 3D structure homology modelling, phylogenetic tree building), the quality which would not be guaranteed in the absence of a recognisable family/domain.

In summary, it should thus be possible to iteratively apply CanOE and BKACE-like studies to gradually cover one or several organism's metabolism, at least when metabolic reactions are already formalised in public metabolic databases.

X.B.3. Reaction dependency

Metabolic factorial analysis is obviously dependent on the metabolic knowledge of the studied organisms and the way it is modelled. Presence/absence of reactions is established by inventorying each genome's reactions, and these are then assigned to metabolic pathways. However, the final goal of metabolic factorial analysis is a very global, exploratory view of the data, and such a high-level representation necessarily distances itself with the underlying data model somewhat.

CanOE is dependent on pre-defined metabolic reactions. Indeed, it works with a global metabolic network that only includes reactions present in the selected metabolic source database. It it thus unable to propose gap reactions corresponding to previously unknown reactions; it is also unable to generate on its own the instances of a given generic reaction (*e.g.* KEGG reaction R07326 describes an alcohol:NAD+ oxidoreductase with a generic alcohol substrate; it has several instances that specify the given alcohol, such as R00754 which is ethanol:NAD+ oxidoreductase). This severely limits the novelty of CanOE propositions, and an interesting perspective would be the development of an automated procedure that might be able to "invent" gap reactions as necessary. Such future works would undoubtedly be inspired by previous works on enzymatic mechanisms [272], metabolic pathway "compound-based" construction [218,221] and compound matching [290,222].

The BKACE strategy has an opposite limitation: it relies on the semi-manual identification of alternative biochemical reactions from a generic mechanism, but is thus limited to a specific group of similar reactions. This means that alternative gene functions (such as those generated by underground metabolism not related to substrate specificity, multifunctional proteins, or "evolutionary leap-frogging" events that drastically change the function of a protein) cannot be proposed by the BKACE strategy without a manual intervention requiring a massive gene-by-gene analysis effort that is not even guaranteed to deliver. Once again, it appears that using BKACE- and CanOE-like strategies in concert, bound together by manual expertise, might help circumnavigate the limitations of both.

As both are designed to generate novel gene-reaction annotations, one may wonder if would be more interesting, given an *exclusive* strategy choice and a target protein to annotate, to either conduct a BKACE-like study of the protein's family, or a CanOE study of the proposed annotations for this protein. This, obviously, depends on both the requirements of each strategy, and on the objectives one has. If a putative function of the protein is already expected, a BKACE-like study is possible. CanOE does not require this kind of previous knowledge. It does, however, require that any target reaction be already defined; furthermore, if no metabolons containing the protein's coding gene is available, it will be powerless. If the conditions of each strategy is met, the objectives determine the choice. Indeed, if an in-depth study with experimental biochemistry is required, BKACE comes out first, more so if the target reaction(s) is(are) not yet modelled in metabolic databases used by CanOE. All in all, CanOE is only useful in a small number of cases, concerning the annotation of proteins of unknown function in already-located metabolons that contain already-defined reactions, and its results remain purely hypothetical. These results can be, however, instrumental in guiding experimental assays and reducing the functional and gene spaces to explore with them.

To summarise, all the projects presented in this manuscript rely on previous metabolic knowledge from public databases: metabolic factorial analysis and CanOE rely on the annotations of a genome with metabolic reactions (that are assigned to pathways in their model), BKACE on at least one metabolic annotation of a protein from a gene family of interest. The BKACE strategy further requires the identification of the chemical mechanism of its target reaction. So far, few bioinformatics resources integrate chemical mechanisms into their data models.

X.B.4. Reaction definition

All the projects described in this thesis depend on the metabolic data models they work with in some way or another. The way the reactions and pathways are formalised influences how reactions can be connected together in a metabolic network, how instances of generic reactions can be generated, how alternate substrates can be imagined... To our knowledge, however, metabolic databases rarely model reactions in a way that accounts for substrate promiscuity and chemical mechanisms. The best efforts that we know of are the MACiE database [272] that stores mechanisms for 321 EC numbers, and KEGG and MetaCyc, which contain generic reactions that can be linked to substrate-specific reaction instances, but without providing full detail of the different relationships nor full freedom for the creation of new instances.

Generic reactions and compound-specific instances of generic reactions are not the only issue with representing metabolic knowledge in databases. One of the most widely-used reaction ontologies, the Enzyme Commission (EC) number scheme, has many drawbacks. To start with, the number of enzymatic reactions described is lagging behind that of other databases such as KEGG and MetaCyc. Also, the generic/specific aspect of a reaction is not addressed in a single, unified way. In practice, many proteins actually end up being annotated with several EC numbers, and not always due to multi-functionality. EC numbers do not describe chemical mechanisms. Finally, the link between sequence similarity, structural similarity and reaction similarity is not straightforward and filled with exceptions at any level of the EC classification scheme [100]. These issues lead to a new issue, that of establishing correspondences between EC numbers and other models of metabolic reactions. Preliminary work with KEGG schemas in CanOE showed that using EC numbers to infer gene-KEGG reaction annotations generated many spurious correspondences that impeded CanOE use. The use of the Gene Ontology (GO) is not a possibility either, as its terms are currently not adapted for describing metabolic reactions.

The Rhea database [http://www.ebi.ac.uk/rhea/] was created in response to several complaints about pre-existing metabolic reaction databases (such as KEGG, MetaCyc, and the list of EC numbers), concerning reaction equation balance, consistent compound references, as well as known biochemical reaction coverage. Rhea reactions can also be combined or split, giving larger freedom in respect to multi-step reactions. The UniPathWay database [213] was designed to contain heavily curated metabolic data and proposes a highly-formalised hierarchical data model, allowing for metabolic pathways to be made up of elementary "chains" of reactions, which would be particularly useful is designing pathway-based main compound edges for a metabolic network based on it.

In summary, existing metabolic reaction and pathway databases are not designed for handling reaction substrate promiscuity and chemical mechanisms at a good coverage, though many recent efforts have innovations that might prove valuable in designing new metabolic networks for CanOE.

X.B.5. Consequences on knowledge and use of metabolism

Using strategies like metabolic factorial analysis, CanOE and BKACE should ensure that the level of general metabolic knowledge continues to increase. Indeed, it will complete our knowledge of the workings of specific metabolic pathways and open up new opportunities for biological experiments. Proposing new metabolic reactions or variants thereof will allow better pathway identification and completion, as well as deeper insight into substrate specificity-bound underground metabolism. Identifying and validating candidate genes for local or global orphans will generate precedence that may be copied to other genes thanks to functional transfer on the basis of detected homology. These new annotations might then be included in new metabolons, allowing the generation of further novel annotation hypotheses for neighbouring genes. As is the case with functional and relational annotation in their entirety, annotation results will feed future annotations, pushing the boundaries of our metabolic knowledge back ever further.

As our knowledge of prokaryote metabolism progresses, so do the opportunities of exploiting it in biotechnology procedures. Indeed, the discovery of new metabolic reaction-coding genes (of either previously orphan enzymes, or little-known metabolic reactions) opens the door to selectively cloning them, which is an often necessary step in producing the enzyme in industrial quantities [231]. Likewise, using a metabolic activity in a synthetic biology application requires the prior identification of its coding gene so that it may be incorporated into the "building blocks" that are at the heart of this emerging discipline.

X.B.6. CanOE & BKACE feed metabolic analyses

The previously-described advance of general knowledge of prokaryote metabolism may prove central to exploring the metabolic capacities of complex microbial ecosystems as in the Carnoulès project. Indeed, the metabolic factorial analyses are built on whatever functional annotations can be made within the partial - and sometimes mixed up - genomes available in metagenomic studies. Better comprehension and knowledge of global prokaryote metabolism, coupled with increased functional annotation transfer opportunities, should help fill in the blanks concerning metagenome metabolism for (supposedly) separate species. This should help inform the factorial analyses, making it more informative and accurate. This in turn would help delimit more precisely where exactly the different organisms interact metabolically, which is obviously one of the key questions of metagenomic studies.

X.B.7. Microme

The Microme project [www.microme.eu] is currently being deployed in Europe in order to assemble a bioinformatics infrastructure dedicated to prokaryote genome analysis and metabolic pathway reconstruction. Metabolic models generated within this project are destined to be exploited in metabolism-based evolutionary analyses and for biotechnological applications. Microme would obviously benefit directly from improved knowledge of prokaryote metabolism brought about by CanOE and BKACE. Metabolic reconstructions often contain metabolic pathways with one or more reactions with missing coding genes; in order to not "break" the reconstruction by creating dead-end metabolites, these reactions are hypothesised to be present in a "gap filling" step. Increased metabolic knowledge would help eliminate the need for such hypotheses.

X.B.8. Integration with experimental validation

Obviously, putative annotations generated by bioinformatics methods - especially when dealing with novel metabolic reactions or little-known families - should be confirmed or invalidated by wet-lab experiments. In the work presented here, we have strived to approach our biochemist collaborators at the Genoscope in view of validating our predictions. Beyond this work, the necessity for direct integration of bioinformatics predictions and experimental testing is now fully recognised, and at least two projects of interest dedicated to this issue are worth noting here.

The Computational Bridge to Experiments (COMBREX) project [www.combrex.org] is a USA-based NIH-funded project that aims to build a hub to which computational prediction methods could be submitted, where they would be run, and resulting predictions would then be screened for reactions and/or gene families of particular interest. These predictions would then be submitted to biochemistry teams (selected amongst many in order to match the prediction type to the team's expertise) for experimental validation, along with a dedicated small grant to fund the experimentation. To my knowledge, COMBREX is currently being tested on a pilot project launched in 2011 [291].

The Enzyme Function Initiative (EFI) [enzymefunction.org] is a smaller USA-based NIH-funded effort that also aims to integrate gene annotation hypothesis generation with biological testing [292]. They are, however, more "biologically orientated" than COMBREX: *a priori* iso-functional families (not isolated genes) may be proposed as candidate catalysts of a given function, by manual expertise and/or by automated prediction methods; activity screening is carried out *in vitro* and *in vivo*; and enzyme/ligand 3D structure is determined when possible. EFI is even younger than

COMBREX and is currently gauging community interest in order to decide whether the effort may be maintained. The objectives of the BKACE strategy are closer to those of EFI while the CanOE strategy objectives are closer to those of COMBREX.

We can hope that both these efforts will last (or even merge) as they will help standardise the way computational predictions are made and dealt with, guaranteeing higher and more consistent quality of gene annotations in public databases. The use of a grant-based incentive for experimental validation of predictions of particular interest should help speed up the discovery process, though one may wonder at how previous annotation prediction providers will deal with the abandonment of their previous experimental collaborators, and reciprocally. I also wonder at whether the fact that both these efforts are USA-based will prove to be a hindrance for scientific teams (bioinformatics and biologists alike) willing to participate in such a endeavour. Perhaps a European alternative might be necessary to federate efforts this side of the Atlantic (and why not an Asian version as well, mirroring the current GenBank/DDBJ/EMBL triumvirate). The Microme project, for example, could be integrated as a part of (or a collaborator of) such a project.

XI. Annexes

XI.A. Spectral clustering

The term “spectral clustering” actually designates a family of clustering algorithms based on a common protocol. The data describing the objects to cluster is first transformed into a weighted, undirected graph using any metric (*e.g.* k-nearest neighbours graph using Euclidean distance). The weighted n*n adjacency matrix of this graph is then transformed into one of the graph's “Laplacian” matrices (there are at least 3 variants of Laplacian matrix, “un-normalised”, “symmetric” and “random walk”). The Laplacian matrix is then decomposed into eigenvectors. The rows of an n*k matrix derived from the first k eigenvectors can then be clustered using a simple k-means algorithm. This clustering is a spectral clustering of the initial data points.

An good introduction to spectral clustering with more detail can be found in [278].

XI.B. Factorial analysis

Factorial analysis designates a large family of statistics methods that aim to extract underlying information from large data sets. This information is gathered in the form of “factors” or “principle components”, which can be viewed as latent variables that summarise in one way or another the data. To put things graphically, the original data points can be imagined as a multi-dimensional cloud of points, through which factorial analysis tries to find the best planes onto which to project them, in order to capture the greatest part of the cloud's variability as possible. It is like trying to find the best angle at which to take a photograph of an object, in order to make the object as recognisable as possible (*e.g.* a lateral photography of a fish).

Multiple Factorial Analysis methods exist due to the numerous data types, analysis objectives, and philosophies. For the latter, two main approaches are opposed: Anglo-saxon (AS-FA) *versus* French (F-FA) factorial analysis. The differences between the two are linked to the prioritisations each give the objectives of the factorial analysis. Indeed, latent information extraction can serve several purposes:

- **summarising:** optimal low-dimensional data representation
- **discovering:** data exploration
- **interpreting:** explain the data

- **modelling:** confirm a mathematical theory of the underlying generative processes

AS-FA is more bent on explaining the data, whereas F-FA is more concerned with representing the data. The fundamental algorithmic difference between them can be explained by this divergence. Indeed, AS-FA allows found factors to be rotated freely in the factorial space, in order to maximise captured variability: they thus maximise explanation. F-FA, however, imposes that all factors be two-by-two orthogonal, ensuring a comprehensive 2D or 3D graphical representation of the original data points projected into the factorial space, and enforcing independence of found factors.

In the works presented in this thesis, I have consistently applied F-FA, as 1) my training is specific to it, and 2) I have generally used FA to either ease clustering or to generate graphical representations, both of which F-FA is good for.

Different F-FA methods exist for treating data of different types, and I shall list here those that I used in my PhD works. **Principal Components Analysis** (PCA) is for use on only quantitative variables (though quantitative and qualitative variables can be used as illustrative variables). **Correspondence Analysis** (CA) analyses the links between the modalities of two qualitative variables. Its corresponding multivariate method is **Multiple Correspondence Analysis** (MCA), and it supports quantitative and qualitative illustrative variables. **Multiple Factorial Analysis** (MFA) is capable of analysing groups of variables; each group can have variables of only one type (quantitative, qualitative), but each group can be of different types. MFA can be powerful for extracting common underlying factors from multiple data sets. **Hierarchical Multiple Factorial Analysis** (HMFA) is an extension of the previous, allowing a hierarchy to be defined on the variable groupings.

All these methods can be found described in detail in [293].

XI.C. Some basics in graph theory

As a support for the notions handled at times in this manuscript, I wanted to present some of the basic concepts from graph theory that could be useful to the reader.

XI.C.1. Graphs and graph elements

A graph is a mathematical structure generally used to model real-life phenomenons in which certain objects present one-to-one (or even many-to-many) relationships. Each object is modelled as a vertex (also known as a node), and relationships are modelled as edges that connect the vertices.

Edges can be binary (*i.e.* either present or absent) or weighted (*i.e.* with a certain real-valued weight that measures the importance of the edge). They can also be undirected (*i.e.* with no specific direction) or directed (*i.e.* the relationship applies specifically from one node to another). In some special types of graph, special types of edges can be found: multiple edges can connect the same pair of vertices, or a same edge can connect multiple vertices (hyperedges), or an edge can connect a vertex to itself (self-edges or self-loops)...

XI.C.2. Matrices of interest

All of the matrices presented here are often used in graph theory-based analyses.

For a graph containing n vertices, its **(weighted) adjacency matrix A** is a n*n matrix where each value $A_{i,j}$ corresponds to the weight of the edge connecting vertex i to vertex j. Undirected graphs have symmetric adjacency matrices, directed graphs may have asymmetric ones. Self-loop edges are obviously present on the matrix diagonal. The adjacency matrix is the most mathematical representation of a graph.

A **vertex degree** is the number of edges that vertex possesses. When edges are directed, in- and out-degrees can be defined. When edges are weighted, degrees can also be weighted. Mathematically, degrees are often represented along the diagonal of the **degree matrix D**, an n*n matrix.

A **distance matrix** is also a n*n matrix which contains, for each vertex pair (i,j), the length of the shortest path between i and j. Paths are defined in the following section.

A graph's **Laplacian matrix** is related to D and A, though its exact matricial definition depends on its type. Indeed, it can be either normalised or not, and the normalised version has several sub-versions.

XI.C.3. Paths, walks, breadth and depth-first searches

A “**path**” is a sequence of vertices in a graph that can be reached by traversing the edges between them, in order. It most of the literature, the definition of “path” (often implicitly) requires that no vertex be traversed more than once; a sequence of vertices where a single vertex can be visited more than once is then referred to as a “**walk**”.

Establishing the shortest path between two vertices is common practice in graph theory as it can give an idea on the topology of the graph. Two families of algorithms are generally used to determine shortest paths. **Depth-first searches**, starting from a first vertex, recursively move through the graph from a given vertex to a neighbour vertex, until a target vertex is reached or a maximum path length has been attained. **Breadth-first searches**, starting from a first vertex, consider all the neighbours of the current vertex; if none of them are the target vertex, then it moves on to one neighbour and proceeds recursively until the target vertex is found or a maximum path length has been attained.

XI.C.4. Graph structures of interest

Given its nature, the most interesting features about a graph generally involve its structure or topology. Vertices in a graph can form **clusters**, which can be described as groups of vertices that are more highly connected amongst themselves than with other vertices. A **hub** is a vertex that is connected to many other vertices which, for themselves, are lowly connected to other vertices, forming a star-like structure; in biology, hubs are generally seen as vertices of central importance. A **clique** is a special cluster, in which all vertices are connected to all other vertices in the cluster. Finally, a **connected component** is a group of vertices in which each vertex can be reached from any other vertex along paths of any length.

This concludes my brief tour of the basics of graph theory.

XII. Bibliography

1. Chubb D, Jefferys BR, Sternberg MJE, Kelley LA (2010) Sequencing delivers diminishing returns for homology detection: implications for mapping the protein universe. Bioinformatics 26: 2664 – 2671. doi:10.1093/bioinformatics/btq527

2. Pouliot Y, Karp PD (2007) A survey of orphan enzyme activities. BMC Bioinformatics 8: 244–244. doi:10.1186/1471-2105-8-244

3. The UniProt Consortium (2010) Ongoing and future developments at the Universal Protein Resource. Nucleic Acids Research 39: D214–D219. doi:10.1093/nar/gkq1020

4. Vallenet D, Engelen S, Mornico D, Cruveiller S, Fleury L, et al. (2009) MicroScope: a platform for microbial genome annotation and comparative genomics. Database 2009: bap021. doi:10.1093/database/bap021

5. Boyer F, Morgat A, Labarre L, Pothier J, Viari A (2005) Syntons, metabolons and interactons: an exact graph-theoretical approach for exploring neighbourhood between genomic and functional data. Bioinformatics 21: 4209–4215. doi:10.1093/bioinformatics/bti711

6. Vallenet D, Labarre L, Rouy Z, Barbe V, Bocs S, et al. (2006) MaGe: a microbial genome annotation system supported by synteny results. Nucl. Acids Res. 34: 53–65. doi:10.1093/nar/gkj406

7. SANGER F, TUPPY H (1951) The amino-acid sequence in the phenylalanyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. Biochem. J. 49: 463–481.

8. SANGER F, TUPPY H (1951) The amino-acid sequence in the phenylalanyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. Biochem. J. 49: 481–490.

9. SANGER F, THOMPSON EOP (1953) The amino-acid sequence in the glycyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. Biochem. J. 53: 353–366.

10. SANGER F, THOMPSON EOP (1953) The amino-acid sequence in the glycyl chain of insulin. II. The investigation of peptides from enzymic hydrolysates. Biochem. J. 53: 366–374.

11. Maxam AM, Gilbert W (1977) A new method for sequencing DNA. Proc. Natl. Acad. Sci. U.S.A. 74: 560–564.

12. WATSON JD, CRICK FH (1953) The structure of DNA. Cold Spring Harb. Symp. Quant. Biol. 18: 123–131.

13. WATSON JD, CRICK FH (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 171: 737–738.

14. WATSON JD, CRICK FH (1953) Genetical implications of the structure of deoxyribonucleic acid. Nature 171: 964–967.

15. Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, et al. (1986) Fluorescence detection in automated DNA sequence analysis. Nature 321: 674–679. doi:10.1038/321674a0

16. Venter J, (other authors hidden for clarity) (2001) The sequence of the human genome. Science (New York, N.Y.) 291: 1304–1351.

17. McPherson J, (other authors hidden for clarity) (2001) A physical map of the human genome. Nature: 934–941.

18. Lander E, (other authors hidden for clarity) (2001) Initial sequencing and analsis of the human genome. Nature 15: 860–921.

19. Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends Genet. 24: 133–141. doi:10.1016/j.tig.2007.12.007

20. McPherson JD (2009) Next-generation gap. Nat Meth 6: S2–S5. doi:10.1038/nmeth.f.268

21. Wolinsky H (2007) The thousand-dollar genome. EMBO Rep 8: 900–903. doi:10.1038/sj.embor.7401070

22. Sboner A, Mu XJ, Greenbaum D, Auerbach RK, Gerstein MB (2011) The real cost of sequencing: higher than you think! Genome Biol 12: 125. doi:10.1186/gb-2011-12-8-125

23. Hugenholtz P (2002) Exploring prokaryotic diversity in the genomic era. Genome Biol 3: REVIEWS0003.

24. Galperin MY, Koonin EV (2010) From complete genome sequence to [`]complete' understanding? Trends in Biotechnology 28: 398–406. doi:16/j.tibtech.2010.05.006

25. Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, et al. (2009) A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. Nature 462: 1056–1060. doi:10.1038/nature08656

26. 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073. doi:10.1038/nature09534

27. Hugenholtz P, Goebel BM, Pace NR (1998) Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. J. Bacteriol 180: 4765–4774.

28. Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chemistry & Biology 5: R245–R249. doi:10.1016/S1074-5521(98)90108-9

29. Wooley JC, Godzik A, Friedberg I (2010) A Primer on Metagenomics. PLoS Comput Biol 6: e1000667. doi:10.1371/journal.pcbi.1000667

30. Savage DC (1977) Microbial ecology of the gastrointestinal tract. Annu. Rev. Microbiol 31: 107–133. doi:10.1146/annurev.mi.31.100177.000543

31. Berg RD (1996) The indigenous gastrointestinal microflora. Trends Microbiol 4: 430–435.

32. Magrane M, UniProt Consortium (2011) UniProt Knowledgebase: a hub of integrated protein data. Database (Oxford) 2011: bar009. doi:10.1093/database/bar009

33. Médigue C, Bocs S, Labarre L, Mathé C, Vallenet D (2002) L'annotation in silico des séquences génomiques - Bio-informatique (1). médecine/sciences 18: 14. doi:10.1051/medsci/2002182237

34. Gaudriault S, Vincent R (2009) Séquençage des Génomes Complets. Génomique. Mémento Sciences. De Boeck Université. p. 37–40.

35. Zhou P, Emmert D, Zhang P (2006) Using Chado to store genome annotation data. Curr Protoc Bioinformatics Chapter 9: Unit 9.6. doi:10.1002/0471250953.bi0906s12

36. Carver T, Berriman M, Tivey A, Patel C, Bohme U, et al. (2008) Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. Bioinformatics 24: 2672–2676. doi:10.1093/bioinformatics/btn529

37. Markowitz VM, Chen I-MA, Palaniappan K, Chu K, Szeto E, et al. (2010) The integrated microbial genomes system: an expanding comparative analysis resource. Nucleic Acids Res 38: D382–D390. doi:10.1093/nar/gkp887

38. Overbeek R, Larsen N, Walunas T, D'Souza M, Pusch G, et al. (2003) The ERGO genome analysis and discovery system. Nucleic Acids Res 31: 164–171.

39. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang H-Y, et al. (2005) The SEED : The Subsystems Approach to Genome Annotation and its Use in the Project to Annotate 1000 Genomes. Nucleic Acids Res 33: 5691–5702. doi:10.1093/nar/gki866

40. Bryson K, Loux V, Bossy R, Nicolas P, Chaillou S, et al. (2006) AGMIAL: implementing an annotation strategy for prokaryote genomes as a distributed system. Nucleic Acids Res 34: 3533–3545. doi:10.1093/nar/gkl471

41. Deonier RC, Tavaré S, Waterman MS (2005) Computational genome analysis: an introduction. Springer. p.

42. Pal D, Eisenberg D (2005) Inference of protein function from protein structure. Structure 13: 121–130. doi:10.1016/j.str.2004.10.015

43. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene Ontology: tool for the unification of biology. Nat Genet 25: 25–29. doi:10.1038/75556

44. Riley M (1993) Functions of the gene products of Escherichia coli. Microbiol. Rev. 57: 862–952.

45. Serres MH, Riley M (2000) MultiFun, a multifunctional classification scheme for Escherichia coli K-12 gene products. Microb. Comp. Genomics 5: 205–222.

46. Ruepp A, Zollner A, Maier D, Albermann K, Hani J, et al. (2004) The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Res 32: 5539–5545. doi:10.1093/nar/gkh894

47. Tanenbaum DM, Goll J, Murphy S, Kumar P, Zafar N, et al. (2010) The JCVI standard operating procedure for annotating prokaryotic metagenomic shotgun sequencing data. Stand Genomic Sci 2: 229–237. doi:10.4056/sigs.651139

48. Haft DH, Selengut JD, Brinkac LM, Zafar N, White O (2005) Genome Properties: a system for the investigation of prokaryotic genetic content for microbiology, genome annotation and comparative genomics. Bioinformatics 21: 293 –306. doi:10.1093/bioinformatics/bti015

49. Magrane M, Consortium U (2011) UniProt Knowledgebase: a hub of integrated protein data. Database (Oxford) 2011. doi:10.1093/database/bar009

50. Remm M, Storm CEV, Sonnhammer ELL (2001) InParanoid : Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. Journal of Molecular Biology 314: 1041–1052. doi:10.1006/jmbi.2000.5197

51. Altenhoff AM, Dessimoz C (2009) Phylogenetic and functional assessment of orthologs inference projects and methods. PLoS Comput. Biol. 5: e1000262. doi:10.1371/journal.pcbi.1000262

52. Bar D (2011) Evidence of Massive Horizontal Gene Transfer Between Humans and Plasmodium vivax. Nature Precedings. Available: http://precedings.nature.com/documents/5690/version/1/. Consulté 27 juill 2011.

53. Schnoes AM, Brown SD, Dodevski I, Babbitt PC (2009) Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies. PLoS Comput Biol 5. doi:10.1371/journal.pcbi.1000605

54. Oliver S (2000) Proteomics: Guilt-by-association goes global. Nature 403: 601–603. doi:10.1038/35001165

55. Janga SC, Díaz-Mejía JJ, Moreno-Hagelsieb G (2011) Network-based function prediction and interactomics: The case for metabolic enzymes. Metabolic Engineering 13: 1–10. doi:16/j.ymben.2010.07.001

56. Sharan R, Ulitsky I, Shamir R (2007) Network-based prediction of protein function. Mol. Syst. Biol 3: 88. doi:10.1038/msb4100129

57. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) BLAST : Basic local alignment search tool. J. Mol. Biol 215: 403–410. doi:10.1006/jmbi.1990.9999

58. Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. U.S.A. 85: 2444–2448.

59. Smith T, Waterman M (1981) Identification of common molecular subsequences. Journal of Molecular Biology 147: 195–197. doi:10.1016/0022-2836(81)90087-5

60. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402.

61. Karplus K, Barrett C, Hughey R (1998) Hidden Markov models for detecting remote protein homologies. Bioinformatics 14: 846–856.

62. Söding J (2005) Protein homology detection by HMM–HMM comparison. Bioinformatics 21: 951 –960. doi:10.1093/bioinformatics/bti125

63. Söding J, Biegert A, Lupas AN (2005) The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 33: W244–W248. doi:10.1093/nar/gki408

64. Tatusov RL, Mushegian AR, Bork P, Brown NP, Hayes WS, et al. (1996) Metabolism and evolution of Haemophilus influenzae deduced from a whole-genome comparison with Escherichia coli. Current Biology 6: 279–291. doi:16/S0960-9822(02)00478-5

65. Andrade MA, Sander C (1997) Bioinformatics: from genome data to biological knowledge. Current Opinion in Biotechnology 8: 675–683. doi:16/S0958-1669(97)80118-8

66. Enright AJ, Van Dongen S, Ouzounis CA (2002) TRIBE-MCL : An efficient algorithm for large-scale detection of protein families. Nucl. Acids Res. 30: 1575–1584. doi:10.1093/nar/30.7.1575

67. van Dongen S (2000) Graph Clustering by Flow Simulation Utrecht. Available: http://www.micans.org/mcl/.

68. Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Research 13: 2178–2189. doi:10.1101/gr.1224503

69. Mavromatis K, Chu K, Ivanova N, Hooper SD, Markowitz VM, et al. (2009) Gene Context Analysis in the Integrated Microbial Genomes (IMG) Data Management System. PLoS ONE 4: e7979. doi:10.1371/journal.pone.0007979

70. Levy E, Ouzounis C, Gilks W, Audit B (2005) Probabilistic annotation of protein sequences based on functional classifications. BMC Bioinformatics 6: 302. doi:10.1186/1471-2105-6-302

71. Audit B, Levy ED, Gilks WR, Goldovsky L, Ouzounis CA (2007) CORRIE: enzyme sequence annotation with confidence estimates. BMC Bioinformatics 8: S3–S3. doi:10.1186/1471-2105-8-S4-S3

72. Tatusov RL, Galperin MY, Natale DA, Koonin EV (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Research 28: 33 –36. doi:10.1093/nar/28.1.33

73. Gattiker A, Michoud K, Rivoire C, Auchincloss AH, Coudert E, et al. (2003) Automated annotation of microbial proteomes in SWISS-PROT. Computational Biology and Chemistry 27: 49–58. doi:10.1016/S1476-9271(02)00094-4

74. Lima T, Auchincloss AH, Coudert E, Keller G, Michoud K, et al. (2009) HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot. Nucleic Acids Research 37: D471–D478. doi:10.1093/nar/gkn661

75. Meyer F, Overbeek R, Rodriguez A (2009) FIGfams: yet another set of protein families. Nucleic Acids Res 37: 6643–6654. doi:10.1093/nar/gkp698

76. Sonnhammer EL, Eddy SR, Durbin R (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins 28: 405–420.

77. Bateman A, Coggill P, Finn RD (2010) DUFs: families in search of function. Acta Crystallogr Sect F Struct Biol Cryst Commun 66: 1148–1152. doi:10.1107/S1744309110001685

78. Claudel-Renard C, Chevalet C, Faraut T, Kahn D (2003) PRIAM : « Enzyme-specific profiles for genome annotation: PRIAM ». Nucl. Acids Res. 31: 6633–6639. doi:10.1093/nar/gkg847

79. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, et al. (2000) InterPro--an integrated documentation resource for protein families, domains and functional sites. Bioinformatics 16: 1145–1150.

80. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, et al. (2009) InterPro: the integrative protein signature database. Nucleic Acids Research 37: D211–D215. doi:10.1093/nar/gkn785

81. Hofmann K, Bucher P, Falquet L, Bairoch A (1999) The PROSITE database, its status in 1999. Nucleic Acids Res. 27: 215–219.

82. Attwood TK, Croning MD, Flower DR, Lewis AP, Mabey JE, et al. (2000) PRINTS-S: the database formerly known as PRINTS. Nucleic Acids Res. 28: 225–227.

83. Henikoff JG, Pietrokovski S, McCallum CM, Henikoff S (2000) Blocks-based methods for detecting protein homology. Electrophoresis 21: 1700–1706. doi:10.1002/(SICI)1522-2683(20000501)21:9<1700::AID-ELPS1700>3.0.CO;2-V

84. Corpet F, Servant F, Gouzy J, Kahn D (2000) ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. Nucleic Acids Res. 28: 267–269.

85. Yeats C, Lees J, Carter P, Sillitoe I, Orengo C (2011) The Gene3D Web Services: a platform for identifying, annotating and comparing structural domains in protein sequences. Nucleic Acids Res. 39: W546–550. doi:10.1093/nar/gkr438

86. Letunic I, Copley RR, Pils B, Pinkert S, Schultz J, et al. (2006) SMART 5: domains in the context of genomes and networks. Nucleic Acids Res. 34: D257–260. doi:10.1093/nar/gkj079

87. Wilson D, Madera M, Vogel C, Chothia C, Gough J (2007) The SUPERFAMILY database in 2007: families and functions. Nucleic Acids Res. 35: D308–313. doi:10.1093/nar/gkl910

88. Nikolskaya AN, Arighi CN, Huang H, Barker WC, Wu CH (2006) PIRSF family classification system for protein functional and evolutionary analysis. Evol. Bioinform. Online 2: 197–209.

89. Haft DH, Selengut JD, White O (2003) The TIGRFAMs database of protein families. Nucleic Acids Res. 31: 371–373.

90. Mi H, Guo N, Kejariwal A, Thomas PD (2007) PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways. Nucleic Acids Res. 35: D247–252. doi:10.1093/nar/gkl869

91. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The Protein Data Bank. Nucleic Acids Research 28: 235 –242. doi:10.1093/nar/28.1.235

92. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, et al. (1997) CATH--a hierarchic classification of protein domain structures. Structure 5: 1093–1108.

93. Greene LH, Lewis TE, Addou S, Cuff A, Dallman T, et al. (2007) The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. Nucleic Acids Res. 35: D291–297. doi:10.1093/nar/gkl959

94. Yeats C, Lees J, Reid A, Kellam P, Martin N, et al. (2008) Gene3D: comprehensive structural and functional annotation of genomes. Nucleic Acids Res. 36: D414–418. doi:10.1093/nar/gkm1019

95. de Melo-Minardi RC, Bastard K, Artiguenave F (2010) Identification of subfamily-specific sites based on active sites modeling and clustering. Bioinformatics 26: 3075–3082. doi:10.1093/bioinformatics/btq595

96. Pierri CL, Parisi G, Porcelli V (2010) Computational approaches for protein function prediction: a combined strategy from multiple sequence alignment to molecular docking-based virtual screening. Biochim. Biophys. Acta 1804: 1695–1712. doi:10.1016/j.bbapap.2010.04.008

97. Shah I, Hunter L (1997) Predicting enzyme function from sequence: a systematic appraisal. Proc Int Conf Intell Syst Mol Biol 5: 276–283.

98. Devos D, Valencia A (2000) Practical limits of function prediction. Proteins 41: 98–107.

99. Gerlt JA, Babbitt PC (2000) Can sequence determine function? Genome Biol 1: REVIEWS0005.

100. Babbitt PC (2003) Definitions of enzyme function for the structural genomics era. Current Opinion in Chemical Biology 7: 230–237. doi:10.1016/S1367-5931(03)00028-0

101. O’Brien PJ, Herschlag D (1999) Catalytic promiscuity and the evolution of new enzymatic activities. Chemistry & Biology 6: R91–R105. doi:10.1016/S1074-5521(99)80033-7

102. James LC, Tawfik DS (2001) Catalytic and binding poly-reactivities shared by two unrelated

proteins: The potential role of promiscuity in enzyme evolution. Protein Sci. 10: 2600–2607. doi:10.1110/ps.14601

103. Copley SD (2003) Enzymes with extra talents: moonlighting functions and catalytic promiscuity. Current Opinion in Chemical Biology 7: 265–272. doi:10.1016/S1367-5931(03)00032-2

104. Bornscheuer UT, Kazlauskas RJ (2004) Catalytic Promiscuity in Biocatalysis: Using Old Enzymes to Form New Bonds and Follow New Pathways. Angewandte Chemie International Edition 43: 6032–6040. doi:10.1002/anie.200460416

105. Hult K, Berglund P (2007) Enzyme promiscuity: mechanism and applications. Trends Biotechnol. 25: 231–238. doi:10.1016/j.tibtech.2007.03.002

106. Nobeli I, Favia AD, Thornton JM (2009) Protein promiscuity and its implications for biotechnology. Nat. Biotechnol 27: 157–167. doi:10.1038/nbt1519

107. Jensen RA (1976) Enzyme recruitment in evolution of new function. Annu. Rev. Microbiol. 30: 409–425. doi:10.1146/annurev.mi.30.100176.002205

108. Tawfik OK and DS (2010) Enzyme Promiscuity: A Mechanistic and Evolutionary Perspective. Annual Review of Biochemistry 79: 471–505. doi:10.1146/annurev-biochem-030409-143718

109. Urbach D, Moore JH (2011) Data mining and the evolution of biological complexity. BioData Mining 4: 7. doi:10.1186/1756-0381-4-7

110. Gibson G, Wagner G (2000) Canalization in evolutionary genetics: a stabilizing theory? Bioessays 22: 372–380. doi:10.1002/(SICI)1521-1878(200004)22:4<372::AID-BIES7>3.0.CO;2-J

111. Bork P, Dandekar T, Diaz-Lazcoz Y, Eisenhaber F, Huynen M, et al. (1998) Predicting function: from genes to genomes and back. J. Mol. Biol 283: 707–725. doi:10.1006/jmbi.1998.2144

112. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, et al. (1999) Detecting protein function and protein-protein interactions from genome sequences. Science 285: 751–753.

113. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D (1999) A combined algorithm for genome-wide prediction of protein function. Nature 402: 83–86. doi:10.1038/47048

114. Kozak M (1983) Comparison of initiation of protein synthesis in procaryotes, eucaryotes, and organelles. Microbiol. Rev 47: 1–45.

115. Koonin EV, Galperin MY (1997) Prokaryotic genomes: the emerging paradigm of genome-based microbiology. Curr. Opin. Genet. Dev. 7: 757–763.

116. Meaburn KJ, Misteli T (2007) Cell biology: Chromosome territories. Nature 445: 379–781. doi:10.1038/445379a

117. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N (1999) Use of contiguity on the chromosome to predict functional coupling. In Silico Biol. (Gedrukt) 1: 93–108.

118. Overbeek R (1999) The use of gene clusters to infer functional coupling. Proceedings of the National Academy of Sciences 96: 2896–2901. doi:10.1073/pnas.96.6.2896

119. Huerta AM, Salgado H, Thieffry D, Collado-Vides J (1998) RegulonDB: a database on transcriptional regulation in Escherichia coli. Nucleic Acids Res 26: 55–59.

120. Chen X, Su Z, Dam P, Palenik B, Xu Y, et al. (2004) Operon prediction by comparative genomics: an application to the Synechococcus sp. WH8102 genome. Nucleic Acids Res. 32: 2147–2157. doi:10.1093/nar/gkh510

121. Price MN, Huang KH, Alm EJ, Arkin AP (2005) A novel method for accurate operon predictions in all sequenced prokaryotes. Nucleic Acids Res. 33: 880–892. doi:10.1093/nar/gki232

122. Okuda S, Katayama T, Kawashima S, Goto S, Kanehisa M (2006) ODB: a database of operons accumulating known operons across multiple genomes. Nucleic Acids Res. 34: D358–362. doi:10.1093/nar/gkj037

123. Mao F, Dam P, Chou J, Olman V, Xu Y (2009) DOOR: a database for prokaryotic operons. Nucleic Acids Res. 37: D459–463. doi:10.1093/nar/gkn757

124. Dehal PS, Joachimiak MP, Price MN, Bates JT, Baumohl JK, et al. (2010) MicrobesOnline: an integrated portal for comparative and functional genomics. Nucleic Acids Res. 38: D396–400. doi:10.1093/nar/gkp919

125. Chuang L-Y, Tsai J-H, Yang C-H (2010) PPO: Predictor for Prokaryotic Operons. Bioinformatics 26: 3127 –3128. doi:10.1093/bioinformatics/btq601

126. Passarge E, Horsthemke B, Farber RA (1999) Incorrect use of the term synteny. Nat Genet 23: 387. doi:10.1038/70486

127. Gilbert DG (2007) DroSpeGe: rapid access database for new Drosophila species genomes. Nucleic Acids Res. 35: D480–485. doi:10.1093/nar/gkl997

128. Novo M, Bigey F, Beyne E, Galeote V, Gavory F, et al. (2009) Eukaryote-to-eukaryote gene transfer events revealed by the genome sequence of the wine yeast Saccharomyces cerevisiae EC1118. Proc. Natl. Acad. Sci. U.S.A. 106: 16333–16338. doi:10.1073/pnas.0904673106

129. Dandekar T, Snel B, Huynen M, Bork P (1998) Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem. Sci 23: 324–328.

130. Yanai I, Mellor JC, DeLisi C (2002) Identifying functional links between genes using conserved chromosomal proximity. Trends Genet. 18: 176–179.

131. Kolesov G, Mewes H-W, Frishman D (2001) SNAPping up functionally related genes based on context information: a colinearity-free approach. Journal of Molecular Biology 311: 639–656. doi:10.1006/jmbi.2001.4701

132. Kolesov G, Mewes H-W, Frishman D (2002) SNAPper: gene order predicts gene function. Bioinformatics 18: 1017–1019.

133. Chen Y, Mao F, Li G, Xu Y (2011) Genome-wide discovery of missing genes in biological pathways of prokaryotes. BMC Bioinformatics 12 Suppl 1: S1. doi:10.1186/1471-2105-12-S1-S1

134. Bowers PM, Pellegrini M, Thompson MJ, Fierro J, Yeates TO, et al. (2004) Prolinks: a database of protein functional linkages derived from coevolution. Genome Biol. 5: R35. doi:10.1186/gb-2004-5-5-r35

135. Lemoine F, Labedan B, Lespinet O (2008) SynteBase/SynteView: a tool to visualize gene order conservation in prokaryotic genomes. BMC Bioinformatics 9: 536. doi:10.1186/1471-2105-9-536

136. Luc N, Risler J-L, Bergeron A, Raffinot M (2003) Gene teams: a new formalization of gene

clusters for comparative genomics. Comput Biol Chem 27: 59–67.

137. Kim S, Choi J-H, Yang J (2005) Gene teams with relaxed proximity constraint. Proc IEEE Comput Syst Bioinform Conf: 44–55.

138. Pasek S, Bergeron A, Risler J-L, Louis A, Ollivier E, et al. (2005) Identification of genomic features using microsyntenies of domains: domain teams. Genome Res. 15: 867–874. doi:10.1101/gr.3638405

139. Vandepoele K (2002) The Automatic Detection of Homologous Regions (ADHoRe) and Its Application to Microcolinearity Between Arabidopsis and Rice. Genome Research 12: 1792–1801. doi:10.1101/gr.400202

140. Simillion C, Janssens K, Sterck L, Van de Peer Y (2008) i-ADHoRe 2.0: an improved tool to detect degenerated genomic homology using genomic profiles. Bioinformatics 24: 127–128. doi:10.1093/bioinformatics/btm449

141. Tesler G (2002) GRIMM: genome rearrangements web server. Bioinformatics 18: 492–493.

142. Sinha AU, Meller J (2007) Cinteny: flexible analysis and visualization of synteny and genome rearrangements in multiple organisms. BMC Bioinformatics 8: 82. doi:10.1186/1471-2105-8-82

143. Fujibuchi W, Ogata H, Matsuda H, Kanehisa M (2000) Automatic detection of conserved gene clusters in multiple genomes by graph comparison and P-quasi grouping. Nucleic Acids Res 28: 4029–4036.

144. Ogata H, Fujibuchi W, Goto S, Kanehisa M (2000) A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. Nucleic Acids Res 28: 4021–4028.

145. Rödelsperger C, Dieterich C (2010) CYNTENATOR: progressive gene order alignment of 17 vertebrate genomes. PLoS ONE 5: e8861. doi:10.1371/journal.pone.0008861

146. Denielou Y-P, Boyer F, Sagot M-F, Viari A (2008) Recovering isofunctional genes: a multiple genomes synteny-based approach. Lille. Available: http://www2.lifl.fr/jobim2008/actes/jobim08-denielou.pdf. Consulté 6 janv 2010.

147. Denielou Y-P (2010) Alignement Multiple de Données Génomiques et Post-Génomiques : Approches Algorithmiques UNiversité de Grenoble. Available: http://hal.archives-ouvertes.fr/tel-00610419/. Consulté 16 déc 2011.

148. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA (1999) Protein interaction maps for complete genomes based on gene fusion events. Nature 402: 86–90. doi:10.1038/47056

149. Snel B, Bork P, Huynen M (2000) Genome evolution: gene fusion versus gene fission. Trends in Genetics 16: 9–11. doi:10.1016/S0168-9525(99)01924-1

150. Enright AJ, Ouzounis CA (2001) Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions. Genome Biol. 2: RESEARCH0034.

151. Suhre K, Claverie J-M (2004) FusionDB: a database for in-depth analysis of prokaryotic gene fusion events. Nucleic Acids Res 32: D273–D276. doi:10.1093/nar/gkh053

152. van Helden J (2003) Regulatory sequence analysis tools. Nucleic Acids Res. 31: 3593–3596.

153. Thomas-Chollier M, Sand O, Turatsinze J-V, Janky R, Defrance M, et al. (2008) RSAT: regulatory

sequence analysis tools. Nucleic Acids Research 36: W119–W127. doi:10.1093/nar/gkn304

154. Thomas-Chollier M, Defrance M, Medina-Rivera A, Sand O, Herrmann C, et al. (2011) RSAT 2011: regulatory sequence analysis tools. Nucleic Acids Res. 39: W86–91. doi:10.1093/nar/gkr377

155. Gelfand MS, Novichkov PS, Novichkova ES, Mironov AA (2000) Comparative analysis of regulatory patterns in bacterial genomes. Brief. Bioinformatics 1: 357–371.

156. Gama-Castro S, Salgado H, Peralta-Gil M, Santos-Zavaleta A, Muñiz-Rascado L, et al. (2011) RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). Nucleic Acids Res 39: D98–105. doi:10.1093/nar/gkq1110

157. Pachkov M, Erb I, Molina N, van Nimwegen E (2007) SwissRegulon: a database of genome-wide annotations of regulatory sites. Nucleic Acids Res. 35: D127–131. doi:10.1093/nar/gkl857

158. Grote A, Klein J, Retter I, Haddad I, Behling S, et al. (2009) PRODORIC (release 2009): a database and tool platform for the analysis of gene regulation in prokaryotes. Nucleic Acids Res. 37: D61–65. doi:10.1093/nar/gkn837

159. Pellegrini M (1999) Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. Proceedings of the National Academy of Sciences 96: 4285–4288. doi:10.1073/pnas.96.8.4285

160. Ruano-Rubio V, Poch O, Thompson JD (2009) Comparison of eukaryotic phylogenetic profiling approaches using species tree aware methods. BMC Bioinformatics 10: 383. doi:10.1186/1471-2105-10-383

161. Huynen M (2000) Predicting Protein Function by Genomic Context: Quantitative Evaluation and Qualitative Inferences. Genome Research 10: 1204–1210. doi:10.1101/gr.10.8.1204

162. von Mering C von, Huynen M, Jaeggi D, Schmidt S, Bork P, et al. (2003) STRING: a database of predicted functional associations between proteins. Nucl. Acids Res. 31: 258–261. doi:10.1093/nar/gkg034

163. Ferrer L, Dale JM, Karp PD (2010) A systematic study of genome context methods: calibration, normalization and combination. BMC Bioinformatics 11: 493. doi:10.1186/1471-2105-11-493

164. Crawford JW, Crawford S (1980) Research in psychiatry: a co-citation analysis. Am J Psychiatry 137: 52–55.

165. Estabrooks CA, Derksen L, Winther C, Lavis JN, Scott SD, et al. (2008) The intellectual structure and substance of the knowledge utilization field: a longitudinal author co-citation analysis, 1945 to 2004. Implement Sci 3: 49. doi:10.1186/1748-5908-3-49

166. Šarić J, Jensen LJ, Ouzounova R, Rojas I, Bork P (2006) Extraction of regulatory gene/protein networks from Medline. Bioinformatics 22: 645 –650. doi:10.1093/bioinformatics/bti597

167. Hur J, Sullivan KA, Pande M, Hong Y, Sima AAF, et al. (2011) The identification of gene expression profiles associated with progression of human diabetic neuropathy. Brain 134: 3222–3235. doi:10.1093/brain/awr228

168. von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, et al. (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucl. Acids

Res. 33: D433–437. doi:10.1093/nar/gki005

169. Fields S, Song O (1989) A novel genetic system to detect protein-protein interactions. Nature 340: 245–246. doi:10.1038/340245a0

170. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, et al. (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 440: 637–643. doi:10.1038/nature04670

171. Cusick ME, Klitgord N, Vidal M, Hill DE (2005) Interactome: gateway into systems biology. Hum. Mol. Genet 14 Spec No. 2: R171–181. doi:10.1093/hmg/ddi335

172. Bandyopadhyay S, Sharan R, Ideker T (2006) Systematic identification of functional orthologs based on protein network comparison. Genome Res. 16: 428–435. doi:10.1101/gr.4526006

173. Singh R, Xu J, Berger B (2008) Global alignment of multiple protein interaction networks with application to functional orthology detection. Proc. Natl. Acad. Sci. U.S.A. 105: 12763–12768. doi:10.1073/pnas.0806627105

174. Liao C-S, Lu K, Baym M, Singh R, Berger B (2009) IsoRankN: spectral methods for global alignment of multiple protein networks. Bioinformatics 25: i253–258. doi:10.1093/bioinformatics/btp203

175. Gerdes S, Edwards R, Kubal M, Fonstein M, Stevens R, et al. (2006) Essential genes on metabolic maps. Current Opinion in Biotechnology 17: 448–456. doi:16/j.copbio.2006.08.006

176. Zhang Y, Thiele I, Weekes D, Li Z, Jaroszewski L, et al. (2009) Three-Dimensional Structural View of the Central Metabolic Network of Thermotoga maritima. Science 325: 1544–1549. doi:10.1126/science.1174671

177. Osterman A, Overbeek R (2003) Missing genes in metabolic pathways: a comparative genomics approach. Current Opinion in Chemical Biology 7: 238–251. doi:10.1016/S1367-5931(03)00027-9

178. Pavlidis P, Weston J, Cai J, Grundy WN (2001) Gene functional classification from heterogeneous data. ACM Press. p. 249–255. Available: http://dl.acm.org/citation.cfm?id=369228. Consulté 18 sept 2011.

179. Deng M, Tu Z, Sun F, Chen T (2004) Mapping gene ontology to proteins based on protein–protein interaction data. Bioinformatics 20: 895 –902. doi:10.1093/bioinformatics/btg500

180. Joshi T, Chen Y, Becker JM, Alexandrov N, Xu D (2004) Genome-scale gene function prediction using multiple sources of high-throughput data in yeast Saccharomyces cerevisiae. OMICS 8: 322–333. doi:10.1089/omi.2004.8.322

181. Chen Y, Xu D (2005) Genome-scale protein function prediction in yeast Saccharomyces cerevisiae through integrating multiple sources of high-throughput data. Pac Symp Biocomput: 471–482.

182. Zhu F, Han LY, Chen X, Lin HH, Ong S, et al. (2008) Homology-free prediction of functional class of proteins and peptides by support vector machines. Curr. Protein Pept. Sci 9: 70–95.

183. Erdin S, Lisewski AM, Lichtarge O (2011) Protein function prediction: towards integration of similarity metrics. Current Opinion in Structural Biology 21: 180–188. doi:10.1016/j.sbi.2011.02.001

184. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, et al. (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res 39: D561–568. doi:10.1093/nar/gkq973

185. Doerks T, von Mering C, Bork P (2004) Functional clues for hypothetical proteins based on genomic context analysis in prokaryotes. Nucleic Acids Research 32: 6321 –6326. doi:10.1093/nar/gkh973

186. Taboada B, Verde C, Merino E (2010) High accuracy operon prediction method based on STRING database scores. Nucleic Acids Res 38: e130. doi:10.1093/nar/gkq254

187. Jiang X, Gold D, Kolaczyk ED (2010) Network-based Auto-probit Modeling for Protein Function Prediction. Biometrics. Available: http://www.ncbi.nlm.nih.gov/pubmed/21133881. Consulté 24 août 2011.

188. Nitsch D, Tranchevent L-C, Thienpont B, Thorrez L, Van Esch H, et al. (2009) Network Analysis of Differential Expression for the Identification of Disease-Causing Genes. PLoS ONE 4: e5526. doi:10.1371/journal.pone.0005526

189. Nitsch D, Goncalves J, Ojeda F, de Moor B, Moreau Y (2010) Candidate gene prioritization by network analysis of differential expression using machine learning approaches. BMC Bioinformatics 11: 460. doi:10.1186/1471-2105-11-460

190. Nitsch D, Tranchevent L-C, Gonçalves JP, Vogt JK, Madeira SC, et al. (2011) PINTA: a web server for network-based gene prioritization from expression data. Nucleic Acids Res 39: W334–W338. doi:10.1093/nar/gkr289

191. White S, Smyth P (2003) Algorithms for estimating relative importance in networks. Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. KDD ’03. New York, NY, USA: ACM. p. 266–275. Available: http://doi.acm.org/10.1145/956750.956782. Consulté 15 déc 2011.

192. Hu L, Huang T, Liu X-J, Cai Y-D (2011) Predicting Protein Phenotypes Based on Protein-Protein Interaction Network. PLoS One 6. doi:10.1371/journal.pone.0017668

193. Gomez A, Cedano J, Amela I, Planas A, Pinol J, et al. (2011) Gene Ontology Function prediction in Mollicutes using Protein-Protein Association Networks. BMC Systems Biology 5: 49. doi:10.1186/1752-0509-5-49

194. Mellor JC, Yanai I, Clodfelter KH, Mintseris J, DeLisi C (2002) Predictome: a database of putative functional links between proteins. Nucleic Acids Res. 30: 306–309.

195. Mostafavi S, Ray D, Warde-Farley D, Grouios C, Morris Q (2008) GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. Genome Biol. 9 Suppl 1: S4. doi:10.1186/gb-2008-9-s1-s4

196. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, et al. (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res. 38: W214–220. doi:10.1093/nar/gkq537

197. Monod J (1942) Recherches sur la croissance des cultures bactériennes

198. NYC JF, MITCHELL HK (1949) The use of isotopic carbon in a study of the metabolism of anthranilic acid in Neurospora. J. Biol. Chem 179: 783–787.

199. Almonacid DE, Yera ER, Mitchell JBO, Babbitt PC (2010) Quantitative Comparison of Catalytic Mechanisms and Overall Reactions in Convergently Evolved Enzymes: Implications for Classification of Enzyme Function. PLoS Comput Biol 6. doi:10.1371/journal.pcbi.1000700

200. Almonacid DE, Babbitt PC (2011) Toward mechanistic classification of enzyme functions. Curr Opin Chem Biol 15: 435–442. doi:10.1016/j.cbpa.2011.03.008

201. Jain E, Bairoch A, Duvaud S, Phan I, Redaschi N, et al. (2009) Infrastructure for the life sciences: design and implementation of the UniProt website. BMC Bioinformatics 10: 136. doi:10.1186/1471-2105-10-136

202. Bairoch A (2000) The ENZYME database in 2000. Nucl. Acids Res. 28: 304–305. doi:10.1093/nar/28.1.304

203. Chang A, Scheer M, Grote A, Schomburg I, Schomburg D (2009) BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009. Nucleic Acids Res 37: D588–D592. doi:10.1093/nar/gkn820

204. Scheer M, Grote A, Chang A, Schomburg I, Munaretto C, et al. (2011) BRENDA, the enzyme information system in 2011. Nucleic Acids Res. 39: D670–676. doi:10.1093/nar/gkq1089

205. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res 38: D355–360. doi:10.1093/nar/gkp896

206. Karp P, Riley M, Paley SM, Pellegrini-Toole A, Krummenacker M (1997) EcoCyc: Enyclopedia of Escherichia coli Genes and Metabolism. Nucleic Acids Research 25: 43–51. doi:10.1093/nar/25.1.43

207. Keseler IM, Collado-Vides J, Santos-Zavaleta A, Peralta-Gil M, Gama-Castro S, et al. (2010) EcoCyc: a comprehensive database of Escherichia coli biology. Nucleic Acids Research 39: D583–D590. doi:10.1093/nar/gkq1143

208. Karp PD, Krummenacker M, Paley S, Wagg J (1999) Integrated pathway-genome databases and their role in drug discovery. Trends in Biotechnology 17: 275–281. doi:16/S0167-7799(99)01316-5

209. Karp PD, Paley SM, Krummenacker M, Latendresse M, Dale JM, et al. (2010) Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. Brief. Bioinformatics 11: 40–79. doi:10.1093/bib/bbp043

210. Vastrik I, D'Eustachio P, Schmidt E, Joshi-Tope G, Gopinath G, et al. (2007) Reactome: a knowledge base of biologic pathways and processes. Genome Biol. 8: R39. doi:10.1186/gb-2007-8-3-r39

211. Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, et al. (2009) Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res. 37: D619–622. doi:10.1093/nar/gkn863

212. Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, et al. (2011) Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res. 39: D691–697. doi:10.1093/nar/gkq1018

213. Morgat A, Coissac E, Coudert E, Axelsen KB, Keller G, et al. (2011) UniPathway: a resource for the exploration and annotation of metabolic pathways. Nucleic Acids Research. Available: http://nar.oxfordjournals.org/content/early/2011/11/17/nar.gkr1023.abstract. Consulté 11 déc

2011.

214. Karp PD, Caspi R (2011) A survey of metabolic databases emphasizing the MetaCyc family. Arch. Toxicol. 85: 1015–1033. doi:10.1007/s00204-011-0705-2

215. Reddy VN, Mavrovouniotis ML, Liebman MN (1993) Petri net representations in metabolic pathways. Proc Int Conf Intell Syst Mol Biol 1: 328–336.

216. Kharchenko P, Chen L, Freund Y, Vitkup D, Church G (2006) Identifying metabolic enzymes with multiple types of association evidence. BMC Bioinformatics 7: 177. doi:10.1186/1471-2105-7-177

217. Faust K, Croes D, van Helden J (2009) Metabolic pathfinding using RPAIR annotation. J. Mol. Biol 388: 390–414. doi:10.1016/j.jmb.2009.03.006

218. Arita M (2000) Metabolic reconstruction using shortest paths. Simulation Practice and Theory 8: 109–125. doi:16/S0928-4869(00)00006-9

219. Arita M (2003) In Silico Atomic Tracing by Substrate-Product Relationships in Escherichia coli Intermediary Metabolism. Genome Research 13: 2455 –2466. doi:10.1101/gr.1212003

220. Arita M (2004) The metabolic world of Escherichia coli is not small. Proc. Natl. Acad. Sci. U.S.A 101: 1543–1547. doi:10.1073/pnas.0306458101

221. Boyer F, Viari A (2003) Ab initio reconstruction of metabolic pathways. Bioinformatics 19 Suppl 2: ii26–34.

222. Rahman SA, Bashton M, Holliday GL, Schrader R, Thornton JM (2009) Small Molecule Subgraph Detector (SMSD) toolkit. J Cheminform 1: 12. doi:10.1186/1758-2946-1-12

223. Mithani A, Preston GM, Hein J (2009) Rahnuma: hypergraph-based tool for metabolic pathway prediction and network comparison. Bioinformatics 25: 1831–1832. doi:10.1093/bioinformatics/btp269

224. Heath AP, Bennett GN, Kavraki LE (2010) Finding metabolic pathways using atom tracking. Bioinformatics 26: 1548–1555. doi:10.1093/bioinformatics/btq223

225. Vieira G, Sabarly V, Bourguignon P-Y, Durot M, Le Fèvre F, et al. (2011) Core and panmetabolism in Escherichia coli. J. Bacteriol 193: 1461–1472. doi:10.1128/JB.01192-10

226. Durot M, Bourguignon P-Y, Schachter V (2009) Genome-scale models of bacterial metabolism: reconstruction and applications. FEMS Microbiol Rev 33: 164–190. doi:10.1111/j.1574-6976.2008.00146.x

227. Faust K, Dupont P, Callut J, van Helden J (2010) Pathway discovery in metabolic networks by subgraph extraction. Bioinformatics 26: 1211–1218. doi:10.1093/bioinformatics/btq105

228. Durot M, Le Fèvre F, de Berardinis V, Kreimeyer A, Vallenet D, et al. (2008) Iterative reconstruction of a global metabolic model of Acinetobacter baylyi ADP1 using high-throughput growth phenotype and gene essentiality data. BMC Syst Biol 2: 85. doi:10.1186/1752-0509-2-85

229. Fiehn O, Barupal DK, Kind T (2011) Extending Biochemical Databases by Metabolomic Surveys. Journal of Biological Chemistry 286: 23637 –23643. doi:10.1074/jbc.R110.173617

230. D’Ari R, Casadesús J (1998) Underground metabolism. Bioessays 20: 181–186. doi:10.1002/(SICI)1521-1878(199802)20:2<181::AID-BIES10>3.0.CO;2-0

231. Zaparucha A, de Berardinis V, Weissenbach J (2011) Biocatalyse, bioconversion et biotechnologie blanche : des outils du vivant pour la chimie. La chimie prépare notre avenir (vol. 2): 43–50.

232. Orth JD, Palsson BØ (2010) Systematizing the generation of missing metabolic knowledge. Biotechnology and Bioengineering 107: 403–412. doi:10.1002/bit.22844

233. Karp PD (2004) Call for an enzyme genomics initiative. Genome Biol 5: 401. doi:10.1186/gb-2004-5-8-401

234. Roberts RJ (2004) Identifying Protein Function—A Call for Community Action. PLoS Biol 2. doi:10.1371/journal.pbio.0020042

235. Lespinet O, Labedan B (2006) Puzzling over orphan enzymes. Cell. Mol. Life Sci 63: 517–523. doi:10.1007/s00018-005-5520-6

236. Lespinet O, Labedan B (2006) Orphan enzymes could be an unexplored reservoir of new drug targets. Drug Discov. Today 11: 300–305. doi:10.1016/j.drudis.2006.02.002

237. Chen L, Vitkup D (2007) Distribution of orphan metabolic activities. Trends Biotechnol 25: 343–348. doi:10.1016/j.tibtech.2007.06.001

238. Lespinet O, Labedan B (2006) ORENZA: a web resource for studying ORphan ENZyme activities. BMC Bioinformatics 7: 436–436. doi:10.1186/1471-2105-7-436

239. Tranchevent L-C, Capdevila FB, Nitsch D, De Moor B, De Causmaecker P, et al. (2011) A guide to web tools to prioritize candidate genes. Brief. Bioinformatics 12: 22–32. doi:10.1093/bib/bbq007

240. Kreimeyer A, Perret A, Lechaplais C, Vallenet D, Médigue C, et al. (2007) Identification of the Last Unknown Genes in the Fermentation Pathway of Lysine. Journal of Biological Chemistry 282: 7191–7197. doi:10.1074/jbc.M609829200

241. Banci L, Bertini I, Ciofi-Baffoni S, Katsari E, Katsaros N, et al. (2005) A copper(I) protein possibly involved in the assembly of CuA center of bacterial cytochrome c oxidase. Proc. Natl. Acad. Sci. U.S.A 102: 3994–3999. doi:10.1073/pnas.0406150102

242. Ramazzina I, Folli C, Secchi A, Berni R, Percudani R (2006) Completing the uric acid degradation pathway through phylogenetic comparison of whole genomes. Nat. Chem. Biol 2: 144–148. doi:10.1038/nchembio768

243. Gaballa A, Newton GL, Antelmann H, Parsonage D, Upton H, et al. (2010) Biosynthesis and functions of bacillithiol, a major low-molecular-weight thiol in Bacilli. Proc. Natl. Acad. Sci. U.S.A 107: 6482–6486. doi:10.1073/pnas.1000928107

244. Yamanishi Y, Mihara H, Osaki M, Muramatsu H, Esaki N, et al. (2007) Prediction of missing enzyme genes in a bacterial metabolic network. FEBS Journal 274: 2262–2273. doi:10.1111/j.1742-4658.2007.05763.x

245. Akaho S (2001) A kernel method for canonical correlation analysis. IN PROCEEDINGS OF THE INTERNATIONAL MEETING OF THE PSYCHOMETRIC SOCIETY (IMPS2001. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.76.4434.

246. Kharchenko P, Vitkup D, Church GM (2004) Filling gaps in a metabolic network using expression information. Bioinformatics 20 Suppl 1: i178–185. doi:10.1093/bioinformatics/bth930

247. Freund Y, Schapire RE (1997) A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting,. Journal of Computer and System Sciences 55: 119–139. doi:06/jcss.1997.1504

248. Green ML, Karp PD (2007) Using genome-context data to identify specific types of functional associations in pathway/genome databases. Bioinformatics 23: i205–211. doi:10.1093/bioinformatics/btm213

249. Green ML, Karp PD (2006) The outcomes of pathway database computations depend on pathway ontology. Nucleic Acids Research 34: 3687 –3697. doi:10.1093/nar/gkl438

250. Yao Z, Ruzzo WL (2006) A Regression-based K nearest neighbor algorithm for gene function prediction from heterogeneous data. BMC Bioinformatics 7: S11–S11. doi:10.1186/1471-2105-7-S1-S11

251. Aghaie A, Lechaplais C, Sirven P, Tricot S, Besnard-Gonnet M, et al. (2008) New insights into the alternative D-glucarate degradation pathway. J. Biol. Chem 283: 15638–15646. doi:10.1074/jbc.M800487200

252. Fonknechten N, Perret A, Perchat N, Tricot S, Lechaplais C, et al. (2009) A conserved gene cluster rules anaerobic oxidative degradation of L-ornithine. J. Bacteriol 191: 3162–3167. doi:10.1128/JB.01777-08

253. Zheng Y, Szustakowski JD, Fortnow L, Roberts RJ, Kasif S (2002) Computational identification of operons in microbial genomes. Genome Res. 12: 1221–1230. doi:10.1101/gr.200602

254. Gai AT, Habib M, Paul C, Raffinot M (2003) Identifying Common Connected Components of Graphs. p. Available: http://hal-lirmm.ccsd.cnrs.fr/docs/00/26/94/40/PDF/D97.PDF. Consulté 15 déc 2011.

255. Habib M, Paul C, Raffinot M (2004) Maximal Common Connected Sets of Interval Graphs. Dans: Sahinalp SC, Muthukrishnan S, Dogrusoz U, éditeurs. Combinatorial Pattern Matching. Berlin, Heidelberg: Springer Berlin Heidelberg, Vol. 3109. p. 359–372. Available: http://www.springerlink.com/content/3q748wergyuf1201/. Consulté 15 déc 2011.

256. Denielou Y-P, Sagot M-F, Boyer F, Viari A (2011) Bacterial syntenies: an exact approach with gene quorum. BMC Bioinformatics 12: 193. doi:10.1186/1471-2105-12-193

257. Chen F, Mackey AJ, Vermunt JK, Roos DS (2007) Assessing Performance of Orthology Detection Strategies Applied to Eukaryotic Genomes. PLoS ONE 2: e383. doi:10.1371/journal.pone.0000383

258. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30: 1575–1584.

259. Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, et al. (2008) The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucl. Acids Res. 36: D623–631. doi:10.1093/nar/gkm900

260. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28: 27–30.

261. van Helden J, Wernisch L, Gilbert D, Wodak SJ (2002) Graph-based analysis of metabolic networks. Ernst Schering Res. Found. Workshop: 245–274.

262. Chou C-H, Chang W-C, Chiu C-M, Huang C-C, Huang H-D (2009) FMM: a web server for metabolic pathway reconstruction and comparative analysis. Nucleic Acids Res 37: W129–134. doi:10.1093/nar/gkp264

263. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, et al. (2006) From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res 34: D354–357. doi:10.1093/nar/gkj102

264. Faust K (2010) Development, assessment and application of bioinformatics tools for the extraction of pathways from metabolic networks Université Libre de Bruxelles. Available: http://theses.ulb.ac.be/ETD-db/collection/available/ULBetd-11172010-110850/unrestricted/Karoline_Faust_thesis.pdf.

265. Codd EF (1998) A relational model of data for large shared data banks. 1970. MD Comput 15: 162–166.

266. Petrey D, Honig B (2009) Is protein classification necessary? Toward alternative approaches to function annotation. Curr. Opin. Struct. Biol 19: 363–368. doi:10.1016/j.sbi.2009.02.001

267. Tian W, Arakaki AK, Skolnick J (2004) EFICAz: a comprehensive approach for accurate genome-scale enzyme function inference. Nucleic Acids Res 32: 6226–6239. doi:10.1093/nar/gkh956

268. Hsiao T-L, Revelles O, Chen L, Sauer U, Vitkup D (2010) Automatic policing of biochemical annotations using genomic correlations. Nat. Chem. Biol 6: 34–40. doi:10.1038/nchembio.266

269. Rost B (2002) Enzyme function less conserved than anticipated. J. Mol. Biol. 318: 595–608. doi:10.1016/S0022-2836(02)00016-5

270. O’Boyle NM, Holliday GL, Almonacid DE, Mitchell JBO (2007) Using reaction mechanism to measure enzyme similarity. J. Mol. Biol 368: 1484–1499. doi:10.1016/j.jmb.2007.02.065

271. Satoh H, Sacher O, Nakata T, Chen L, Gasteiger J, et al. (1998) Classification of Organic Reactions: Similarity of Reactions Based on Changes in the Electronic Features of Oxygen Atoms at the Reaction Sites1. J. Chem. Inf. Comput. Sci. 38: 210–219. doi:doi: 10.1021/ci9701190

272. Holliday GL, Almonacid DE, Bartlett GJ, O’Boyle NM, Torrance JW, et al. (2007) MACiE (Mechanism, Annotation and Classification in Enzymes): novel tools for searching catalytic mechanisms. Nucl. Acids Res. 35: D515–520. doi:10.1093/nar/gkl774

273. Bellinzoni M, Bastard K, Perret A, Zaparucha A, Perchat N, et al. (2011) 3-Keto-5-aminohexanoate cleavage enzyme: a common fold for an uncommon Claisen-type condensation. J. Biol. Chem. 286: 27399–27405. doi:10.1074/jbc.M111.253260

274. Higgins DG, Sharp PM (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene 73: 237–244.

275. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl. Acids Res. 32: 1792–1797. doi:10.1093/nar/gkh340

276. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, et al. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 7: 539. doi:10.1038/msb.2011.75

277. Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple

sequence alignment based on fast Fourier transform. Nucl. Acids Res. 30: 3059–3066. doi:10.1093/nar/gkf436

278. von Luxburg U (2007) A tutorial on spectral clustering. Statistics and Computing 17: 395–416. doi:10.1007/s11222-007-9033-z

279. Howe K, Bateman A, Durbin R (2002) QuickTree: building huge Neighbour-Joining trees of protein sequences. Bioinformatics 18: 1546–1547.

280. Brown DP, Krishnamurthy N, Sjölander K (2007) SCI-PHY: Automated Protein Subfamily Identification and Classification. PLoS Comput Biol 3: e160. doi:10.1371/journal.pcbi.0030160

281. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, et al. (2007) Comparative protein structure modeling using MODELLER. Curr Protoc Protein Sci Chapter 2: Unit 2.9. doi:10.1002/0471140864.ps0209s50

282. Le Guilloux V, Schmidtke P, Tuffery P (2009) Fpocket: an open source platform for ligand pocket detection. BMC Bioinformatics 10: 168. doi:10.1186/1471-2105-10-168

283. Shatsky M, Nussinov R, Wolfson HJ (2004) A method for simultaneous alignment of multiple protein structures. Proteins 56: 143–156. doi:10.1002/prot.10628

284. Fisher DH (1987) COBWEB:Knowledge acquisition via incremental conceptual clustering. Machine Learning - 25th anniversary.Vol. 2. p. 139–172. Available: http://www.springerlink.com/content/x8552ppn35245112/. Consulté 30 avr 2010.

285. Holmes G, Donkin A, Witten IH (1994) WEKA: a machine learning workbench. Proceedings of the 1994 Second Australian and New Zealand Conference on Intelligent Information Systems,1994. IEEE. p. 357–361. doi:10.1109/ANZIIS.1994.396988

286. Pei J, Cai W, Kinch LN, Grishin NV (2006) Prediction of functional specificity determinants from protein sequences using log-likelihood ratios. Bioinformatics 22: 164 –171. doi:10.1093/bioinformatics/bti766

287. Strehl A, Ghosh J (2003) Cluster ensembles --- a knowledge reuse framework for combining multiple partitions. J. Mach. Learn. Res. 3: 583–617.

288. Hornik K (2005) A CLUE for Cluster Ensembles.. Available: http://www.jstatsoft.org/v14/i12/. Consulté 2 août 2010.

289. Snel B, Lehmann G, Bork P, Huynen MA (2000) STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucl. Acids Res. 28: 3442–3444. doi:10.1093/nar/28.18.3442

290. Hattori M, Okuno Y, Goto S, Kanehisa M (2003) Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. J. Am. Chem. Soc 125: 11853–11865. doi:10.1021/ja036030u

291. Roberts RJ, Chang Y-C, Hu Z, Rachlin JN, Anton BP, et al. (2010) COMBREX: a project to accelerate the functional annotation of prokaryotic genomes. Nucleic Acids Research 39: D11–D14. doi:10.1093/nar/gkq1168

292. Gerlt JA, Allen KN, Almo SC, Armstrong RN, Babbitt PC, et al. (2011) The Enzyme Function Initiative. Biochemistry. Available: http://dx.doi.org/10.1021/bi201312u.

293. Escofier B, Pagès J (2008) Analyses factorielles simples et multiples : Objectifs, méthodes et interprétation. 4e éd. Dunod. p.