

# ISYMOD: a Knowledge Base for Integrated System Modeling

Julie Chabalier<sup>1,2</sup>, Yves Quentin<sup>1</sup>, Cécile Capponi<sup>2</sup> and Gwennaele Fichant<sup>1</sup>

<sup>1</sup> Laboratoire de Chimie Bactérienne, CNRS, 31 chemin Joseph Aiguier, 13402 Marseille cedex 20 - France  
{chabalie,quentin,fichant}@ibsm.cnrs-mrs.fr

<sup>2</sup> Laboratoire d'Informatique Fondamentale de Marseille, 39 rue Joliot Curie 13453 Marseille cedex 13 - France  
capponi@cmi.univ-mrs.fr

**Keywords.** Ontology, Task Modeling, Knowledge Base, Integrated System

## Introduction

Complex biological functions emerge from interactions between proteins in stable supra-molecular assemblies and/or throughout transitory contacts. Most of the time, protein partners of the assemblies are composed of one or several domains which exhibit different biochemical functions. As an illustration, ABC transporter systems, that are involved in import or export of diverse molecules throughout the cytoplasmic membrane, are composed of two ATPase proteins (or domains) hydrolyzing ATP, and two membrane proteins (or domains) forming the membrane channel [1]. Thus, the study of cellular processes requires the identification of different functional units and their integration in an interaction pathway; such complexes are referred as *integrated systems*. The integrated systems that are involved in the exchanges between the bacterial cell and its environment, are of particular interest for biological reasons but also for illustrating computational challenges we are facing. Indeed, these systems are important for the adaptation of the bacteria to its biotope, so the genomic comparative analysis of their repertoires should help to understand the molecular mechanisms that are involved in the adaptation processes of bacterial genomes. In addition, some of them, such as ABC transporters and two component systems, are functionally linked, allowing an extension of the modeling to a higher level. From the computational point of view, both systems are encoded by families of paralogous genes which are among the most numerous in bacterial genomes, and both show the same identification and reconstruction problems. In bacteria, they are composed of different domains encoded by separated genes, some of them having a fuzzy sequence conservation.

In order to establish, in a complete genome, the repertoire of a given integrated system, we have to go further than the first level of genome annotation (gene and functional predictions). This higher level includes the following steps: i) identifying the different partners using different bio-informatic methods according to their sequence properties, ii) reconstructing the systems using assembly rules, and iii) classifying the system into the correct functional subfamily. Information on interaction pathway is not directly accessible from the analysis of the complete sequence. However, this knowledge can be inferred, either by the analysis of the genomic context of the genes involved in such systems, or throughout phylogenetic inferences drawn from multiple genomic comparisons. Therefore, new computational approaches are needed to handle the analysis and modeling of integrated systems. As a first step in this direction, we have implemented a general automated bio-informatic strategy [2]. Its validation has been first done on the ABC transporters, then extended to the two component systems. The strategies rely upon the use of rules and parameters updated with the incoming data. Thus, predictions obtained in initial states can be updated in the upcoming runs with the possibility of incorrectly propagating the modifications in the complex network of dependencies. Such a pitfall can be avoided if the knowledge of the biological objects and the strategy used to predict their characterisation are all embedded in the same environment as it occurs in a knowledge representation system.

## ISYMOD : a knowledge base developed under AROM

ISYMOD attempts to integrate in the same environment, the knowledge, both factual and methodological, concerning the biological integrated systems. It integrates the modeling of two systems (ABC transporters and two components systems), their functional relations, and the modeling of the methods used for their identification and reconstruction (Fig. 1). The modeling and storage of the data are achieved through the development of a domain knowledge base (Dkb), while the modeling of the methods is embedded in a methodological knowledge base (Mkb). Among its originalities, AROM permits the joint existence of classes and  $n$ -ary associations linking them, with a specialization relation among classes but also among associations [3]. The ontology of ISYMOD exploits such new constructs for modeling the physical interactions between proteins forming a functional complex, as well as to represent a functional relation between two integrated systems.

The developed bio-informatics strategies for identifying and reconstructing integrated systems have been declaratively modeled as task organization in the Mkb. A task is an encapsulated method whose input and output result in domain knowledge base objects or tuples. A task may correspond to a biological problem (for example, comparison of two proteins), thus it will be associated to several bio-informatics solving methods (for example, Blast, Fasta, etc.). A strategy comprises a specified task layout and facilities for the selection of the appropriate solving method depending on the context of execution. This is achieved using AROMTASKS, a module coupled to AROM. At each step of the strategy, in order to keep the coherence of the Dkb, only the root classes are instantiated. Indeed, knowledge is increasing along with the bio-informatics strategy and the tuples linking the objects can contain information enhancing the knowledge on the object itself. For example, a family linked to a system can specify the system type. Therefore, before specializing the objects in the class hierarchy we must obtain all the information about them. Then, the last problem to be solved in the identification strategy is the classification of instances of classes and associations. It is achieved by applying AROM's classification extended with a recursive propagation algorithm that allows i) to automatically attach one object, as well as its partners within a tuple (instance of an association), to the most specialized class they can belong to, and ii) to automatically classify the concerned tuples in the right sub-association [4].

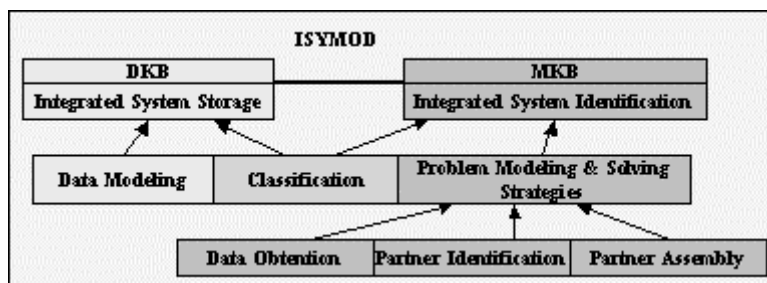


Fig. 1 Overview of ISYMOD architecture

### References

- [1] C.F. Higgins, ABC Transporters: From Microorganisms to Man. *Ann. Rev. Cell Biology*, 8:67-113, 1992
- [2] Y. Quentin, J. Chabalier, J. and G. Fichant, Strategies for the identification, the assembly and the classification of integrated biological systems in completely sequenced genomes. *Computers and Chemistry*, 26: 447-457, 2002.
- [3] C. Capponi, J. Chabalier, Y. Quentin and G. Fichant, A Knowledge Base for Integrated Biological Systems. *IEEE Intelligent Systems*, 16: 52-60, 2001.
- [4] J. Chabalier, G. Fichant and C. Capponi, La classification récursive dans AROM. Application à l'identification de systèmes biologiques. *Actes de la 9<sup>ème</sup> Conférence Francophone Langages et Modèles Objets (LMO'03), Revue des Sciences et Technologies de l'Information (RSTI), série l'Objet*, 9: 167-181, 2003.