GROOLS: Reactive Graph Reasoning for Genome Annotation

Jonathan Mercier[123] and David Vallenet[123]

[1] Direction des Sciences du Vivant, CEA, Institut de Génomique, Genoscope, LABGeM, Evry, France
[2] CNRS-UMR8030, Evry, France
[3] Université d'Evry Val-d'Essonne, Evry, France
{jmercier,vallenet}@genoscope.cns.fr

Abstract. GROOLS (Genomic Rule Oriented Object Logic System) is an expert system to help biologists in the evaluation of genome functional annotation through biological processes like metabolic pathways. Such reasoning is conducted using a Business Rule Management System (BRMS) working on a generic representation of biological knowledge that captures complex data and relationships. We use the Object-Oriented First Order Logic (OOFOL) extended to a four-valued logic of Belnap. Prior biological knowledge is organized in a directed acyclic graph and evaluated by applying reactive graph reasoning using observation facts. Two types of observations are considered: predictions from bioinformatics methods and assertions that correspond to experimental evidences in the studied organism. Once all facts are spread over the graph, a conclusion is made about the state of prior knowledge (*e.g.* confirmed presence, missing, unexpected absence). GROOLS implementation is based on the jBoss DROOLS framework.

Keywords: genome annotation, metabolic network reconstruction, business rules, object-oriented logic, four-valued logic

1 Introduction

GROOLS (Genomic Rule Oriented Object Logic System) is an applied research project mixing biology, informatics and logic. It aims to standardize the use of biological results to annotate a genome. This expert system will help biologists (*i.e.* bio-annotators) to evaluate the quality of gene functional annotation through biological processes like metabolic pathways. Starting from a set of predicted genes, the bio-annotator tries to assign precise molecular functions to corresponding proteins by integrating various predictions from bioinformatics methods, which are mainly based on comparative sequence analysis with proteins having experimentally validated functions. This laborious task may lead to inconsistencies in the annotations due to the lack of experimental results and the difficulty to find a correct trade-off between sensitivity and specificity of the methods for functional inference.

Observations on the organism, like Biolog growth phenotype experiments [5], may help genome annotation notably for metabolic network reconstruction [3]. For example, if an organism is able to grow using a metabolite X then a metabolic pathway for compound X degradation is required in the organism as well as enzymes that catalyze the chemical reactions of the pathway. During the genome annotation process, the consistency and completeness of the predicted enzymatic functions could then be evaluated using this experimental results.
Such reasoning can be conducted using a Business Rule Management System (BRMS) working on a generic representation of biological knowledge that captures complex data and relations like "is-a" and "has-a". These prior knowledge objects are theories and are organized in a directed acyclic graph. In GROOLS, we use the Object-Oriented First Order Logic (OOFOL) [1] extended to a four-valued logic of Belnap [4].

2 Approach

2.1 Metabolic pathway representation

Metabolic pathways are groups of chemical reactions taking part in a same biological process. As depicted in Figure 1, different pathway variants may occur in organisms to achieve a same global chemical transformation. These variants may have common or specific parts, called here reaction blocks.

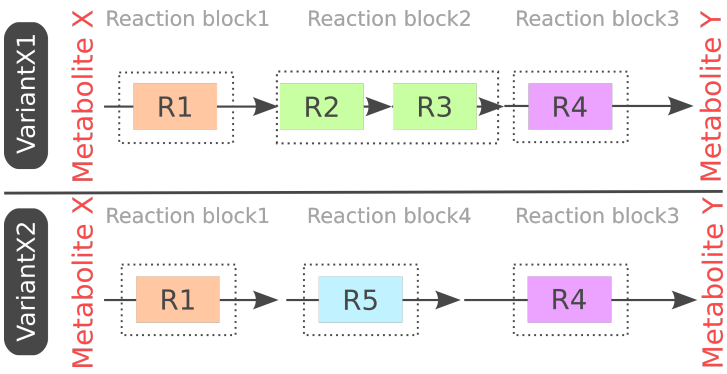

Fig. 1: A metabolic pathway with two variants to transform metabolite X to metabolite Y. These variants share two reaction blocks and are made of 4 reactions for VariantX1 and 3 reactions for VariantX2.

In GROOLS, metabolic pathways with their variants made of reaction blocks will be represented as prior knowledge objects organized in a Directed Acyclic Graph (DAG) (Figure 2). Except for leaf nodes, nodes are flagged with (i) "And"

when a knowledge requires all its sub-knowledge to be present (ii) "Or" when a knowledge requires only one of its sub-knowledge to be present. Considering the example depicted in Figure 2, if the reactions 1, 3 and 4 are present (*i.e.* predicted by gene functional annotation) and the pathwayX is required (*i.e.* the organism degrades compound X) then VariantX1 can be considered as present with one missing reaction (reaction 2). Indeed, biological knowledge is often incomplete and such missing information is commonly called a "pathway hole" or an "orphan enzyme": an enzymatic reaction that should occur in an organism but has no annotated gene to code the enzyme [8].

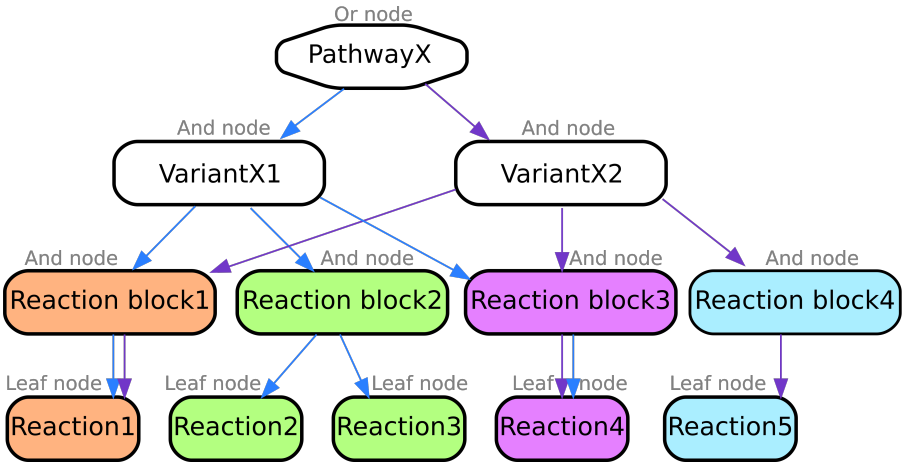

Fig. 2: Directed Acyclic Graph representation of a metabolic pathway.

2.2 Prior knowledge and observation model

A bio-annotator uses at least three types of facts: (i) "prediction facts" are predictions from bioinformatics methods or human expertise made by integrating several method results (ii) "assertion facts" are experimental evidences in the studied organism (iii) "prior knowledge" gathers all metabolic pathways that were experimentally elucidated in at least one organism and represents the actual knowledge over years of cumulative empirical research in biology. Prediction and assertion facts will be considered as observations and will be used to make conclusions about the state of prior biological knowledge. As shown in Figure 3, we designed an object-oriented model with interfaces to ease the integration of heterogeneous objects from different external databases.

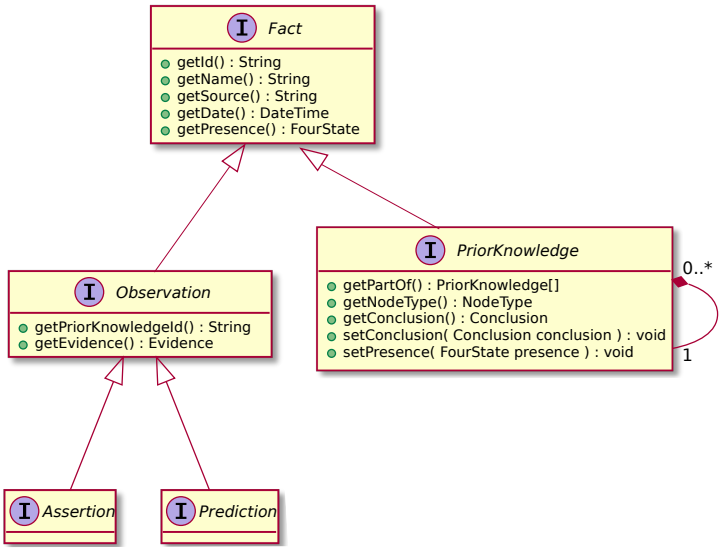

Fig. 3: Grools object-oriented model. Facts are Observations or Prior knowledge. The Prior knowledge is structured in a DAG. Two types of Observations are stored: bioinformatics Predictions or Assertions that are experimental observations.

2.3 Reasoning over structured data

Prior knowledge will be evaluated through observations by applying reactive graph reasoning using the Object-Oriented First Order Logic (OOFOL) [1] extended to a four-valued logic of Belnap [4]. Predictions will directly impact leaf nodes of the DAG while assertions will generally impact root nodes. Prediction and assertions facts take four different values {present, absent, both, unknown} and {required, avoided, both, unknown} respectively, which correspond to {true, false, both, none} values of the four-valued logic. In a logic point of view, a class is a structure and an instance of a class (an object) is a theory. An object can contain other objects meaning that a theory is a graph of smaller theories. Each theory is defined by a coherent set of values. A class C is defined as a tuple $\langle \mathcal{L}_C, \mathcal{A}_C, \mathcal{I}_C \rangle$. Where $\mathcal{L}$ is the vocabulary (class methods), $\mathcal{A}$ a set of axioms (attributes) and $\mathcal{I}$ a set of vocabulary to implement ($\mathcal{I}_C \subset \mathcal{L}_C$). The truth of a theory can be inferred using its sub-theories following a logic. This logic is designed to cope with various and contradictory information sources. A theory is true if all sub-theories are true. If all sub-theories are false then the theory is false. If some sub-theories are true and other false then the theory takes the value "both" to denote ambiguity. If any information matches a theory then the theory takes the state "none" (see Table 1). More exactly, a theory linked to sub-theories with a logical "And" (NodeType knowledge attribute) is evaluated using the following priority "false > both > none > true". For a logical "Or", the

priority is "true > none > both > false". We are also evaluating an optimistic version of the logic where "And" priority is modified to "true > false > both > none" when sub-theories are linked to a single theory. This will help us to deal with incomplete predictions. Lastly, once all facts are spread over the DAG, a conclusion with 16 possible states is made on prior knowledge using assertion and prediction value combination (see Table 2).

Table 1: A four-valued logic truth tables. T: true, F: false, B: both, N: none.

$F(\neg\alpha)$	
T	F
B	B
N	N
F	T

$F(\wedge\alpha)$	**T**	**B**	**N**	**F**
T	T	B	N	F
B	B	B	F	F
N	N	F	N	F
F	F	F	F	F

$F(\vee\alpha)$	**T**	**B**	**N**	**F**
T	T	T	T	T
B	T	B	T	B
N	T	T	N	N
F	T	B	N	F

Table 2: Conclusion truth table.

Assertion \ **Prediction**	**PRESENT**	**ABSENT**	**BOTH**	**UNKNOWN**
REQUIRED	Confirmed P.	Unexpected A.	Contradictory A.	Missing
AVOIDED	Unexpected P.	Confirmed A.	Contradictory P.	Absent
BOTH	Ambiguous P.	Ambiguous A.	Ambiguous C.	Ambiguous
UNKNOWN	Unconfirmed P.	Unconfirmed A.	Unconfirmed C.	Unknown

Legend

A.	Absence
P.	Presence
C.	Contradiction

3 Conclusion and future work

The GROOLS system is still under heavy development and evaluation. We hope that this tool will be useful for biologists to evaluate the overall coherence of individual predicted functions through the integration of additional information from biological processes like metabolic pathways.
Through a collaborative project between INRIA and the Swiss Institute of Bioinformatics, a first prototype with similar deductive reasoning has been implemented in the HERBS system using Jess rule engine. GROOLS implementation is based on the jBoss DROOLS framework which is an open rule-engine written in Java [7]. It natively supports an object-oriented language and uses the PHREAK algorithm to reason over structured data. It does a bridge between our Java application and the business logic. Genomic data with functional predictions and human expert annotations will be extracted from the Prokaryotic

Genome DataBase (PkGDB) of the MicroScope platform[9]. Metabolic pathways will be extracted from MetaCyc [2] and UniPathway [6] resources.

References

1. Amir, E.: Object-oriented first-order logic. Linkiping Electronic Articles in Computer and Information Science (http://www.ida.liu.se/ext/etai) 4, 63–84 (1999), http://www.ep.liu.se/ej/etai/1999/008/
2. Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C.A., Holland, T.A., Keseler, I.M., Kothari, A., Kubo, A., et al.: The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. Nucleic Acids Res 42(D1), D459–D471 (2014)
3. Francke, C., Siezen, R.J., Teusink, B.: Reconstructing the metabolic network of a bacterium from its genome. Trends in microbiology 13(11), 550–558 (2005)
4. Jr, N.D.: A useful four-valued logic. In: Modern uses of multiple-valued logic, pp. 5–37. Springer (1977)
5. Mackie, A.M., Hassan, K.A., Paulsen, I.T., Tetu, S.G.: Biolog phenotype microarrays for phenotypic characterization of microbial cells. In: Environmental Microbiology, pp. 123–130. Springer (2014)
6. Morgat, A., Coissac, E., Coudert, E., Axelsen, K.B., Keller, G., Bairoch, A., Bridge, A., Bougueleret, L., Xenarios, I., Viari, A.: Unipathway: a resource for the exploration and annotation of metabolic pathways. Nucleic Acids Res. p. gkr1023 (2011)
7. Proctor, M., Neale, M., Lin, P., Frandsen, M.: Drools documentation. JBoss.org, Tech. Rep (2008)
8. Sorokina, M., Stam, M., Médigue, C., Lespinet, O., Vallenet, D.: Profiling the orphan enzymes. Biol. Direct 9(10) (2014)
9. Vallenet, D., Belda, E., Calteau, A., Cruveiller, S., Engelen, S., Lajus, A., Le Fevre, F., Longin, C., Mornico, D., Roche, D., Rouy, Z., Salvignol, G., Scarpelli, C., Thil Smith, A.A., Weiman, M., Medigue, C.: MicroScope–an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data. Nucleic Acids Res. 41(Database issue), D636–647 (2013)