Reactive Graph Reasoning for Genomic Annotation

Jonathan Mercier[13], David Vallenet[1], and Claudine Medigue[12]

[1] Direction des Sciences du Vivant, CEA, Institut de Génomique, Genoscope, France
[2] CNRS-UMR8030, Evry, France
[3] Université d'Evry Val d'Essonne, Evry, France

Abstract. The emergence of the next generation sequencing generates an incredible amount of genomes, whereas curation efforts to annotate them tend to decrease despite some community initiatives [5]. Providing human expert annotation as gold standard for all of these genomes became challenging.
To ease this manual process, we develop a Genomic Rule Oriented Object Logic System named GROOLS. The combined use of modern paradigm as oriented object programming with the logic fields provide an interesting space of research. Complex biological information are pulled from various resources and represented as a structured object. These objects are bivalent, for a computer specialist is an instance of a defined class, while for a logician is a theory.
Business Rule is used to coerce, control and making business decision. It separates clearly the business logic from the application. Theses rules are applied to facts throw a Business Rule Management System (BRMS). It evaluates facts in a reproducible and efficient way. This paper show a standardized way to analyze and annotate genomes using a four valued logic and a hierarchical knowledge representation.

Keywords: genomic annotation, network reconstruction, business rules, oriented object logic, four valued logic

1 General description

GROOLS is an applied research project mixing biology, informatics and logic. It aim to standardize the use of biological results to annotate a genome. Highlighting bio-annotator on expected gene or not and discovery new knowledge. During analysis of a newly sequenced genome, the bio-annotator is faced to the difficult task of checking the annotation consistency. In such cases the bio-annotator should to check the proteome against the biological knowledge about the organisms under study. However this task is too huge, bio-annotator usually check sequence similarity to an already annotated protein to infer the gene annotation. The early stages are to understand the bio-annotator reasoning and the identification of a generic model establishing a vocabularies.
A bio-annotator use at least three kinds of facts. The first one is a prediction

fact, found by bioinformatics computing or bio-annotator analysis. The second one is an asserted fact, an empirical evidence on organisms. The thirds is a trusted knowledge, a knowledge known to occur on one ore more organisms. A bio-annotator knowning for the studied organisms that is a tryptophan autotroph. This human expert will expect to found tryptophan bio-synthesis pathway. The corresponding knowledge will be marked as required. Reciprocally, if is known to not grow on tryptophan free media, the tryptophan bio-synthesis pathway is avoided.
Such reasoning system need a generic representation of the biological knowledge to be used on genomic annotation. And, on the other hand a BRMS able to reason over complex and structured data.
In order to have this kind of reasoning we need to select resources, where facts come from. An asserted fact can come from biolog data [3], a trusted knowledge from unipathway [6] and prediction from bioinformatics results or bio-annotator expertise from the Prokaryotic Genome DataBase (PkGDB) [7]
Trusted knowledge are linked together to form an graph. Prediction will influence the leaf of graph while assertion influence the roots of graph. The predictions influences are propagate from the leaf to the roots. Assertions and predictions facts are four valued. An assertion can take one of these values: present, absent, both and unknown and a prediction: required, avoided, both, unknown. These state are inferred to knowledge giving ability to get a precise conclusion. This aims to help bio-annotator to focus on missing gene annotation using a reconstruction of the metabolic network from a sequenced genome [4].
This kind of logic is described by Belnap [2] as a four valued logic. These theories are connected via an interface vocabularies. A set of rules is applied on a huge amount of biological knowledge to notify the completeness and the consistencies for each genome. To achieve this we have applied the Object-Oriented First Order Logic (OOFOL) [1] and extended to a four valued logic. The four valued logic give a finer analysis than classic First Order Logic. In GROOLS, knowledge objects are theories which takes the form of a directed acyclic graph. Others are simple typed fact describing what exists that means all computed prediction or human prediction. Or the others kind of fact what should to exists, an assertion to a theory.

2 Genomic annotation

Genomic annotation task is one of most important, at ends organisms molecular physiology are decoded increasing our understanding of that particular organism. This field aims to understand what an organisms is able to do from his DNA sequence. The unit in genome annotation process is the product of gene (protein) and understanding its function. Some of these proteins are defined as enzymes, that means they are able to catalyze at least one chemical reaction. The enzymatic product of theses chemicals reactions will be used by an enzyme and so on. This enzyme chained process is described by a metabolic pathway (Figure 1). An

enzyme can to be depicted in more than one metabolism pathway to form a metabolic network.

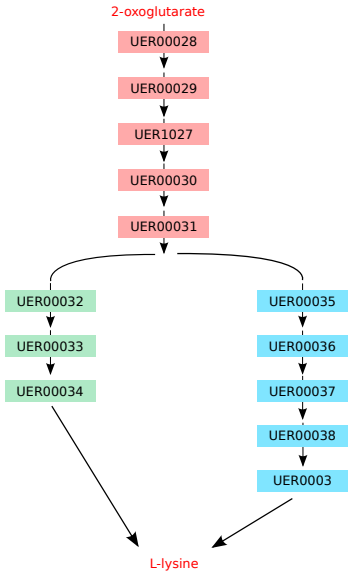

Fig. 1: A metabolic pathway with two path to go from 2-oxoglutarate metabolite to L-Lysine . UER means unipathway enzymatic reaction which are involved by an enzymatic protein. Box with a same color are in a same block of reaction. This pathway has three reactions block: red green and blue. An organism could have a path with red and green blocks and/or red and blue block. Source: unipathway .

A bio-annotator usually defined a coding region, search an homologous sequence and take a look to the genomic context to infer an annotation. In our opinion this task is more productive to do after an automatic genomic annotation. This allow the reconstruction of metabolic process (see [4]), filled by automatics tasks, and help the bio-annotator to focus on missing gene. This automatics gene annotation information can to be displayed with some others information enabling a better understanding on the organism abilities (depicted in Figure 2).

This kind of reasoning should to be done over more than four thousand genes by organisms. Moreover genome to annotate come quickly than bio-annotator community can do. On the other hand human expertise can lead to some inconsistencies which will be reused by others bio-annotator. To avoid these issues we develop an expert system to automatize the bio-annotator reasoning. An automated reasoner can manage billions of facts in a standard and reproducible way.

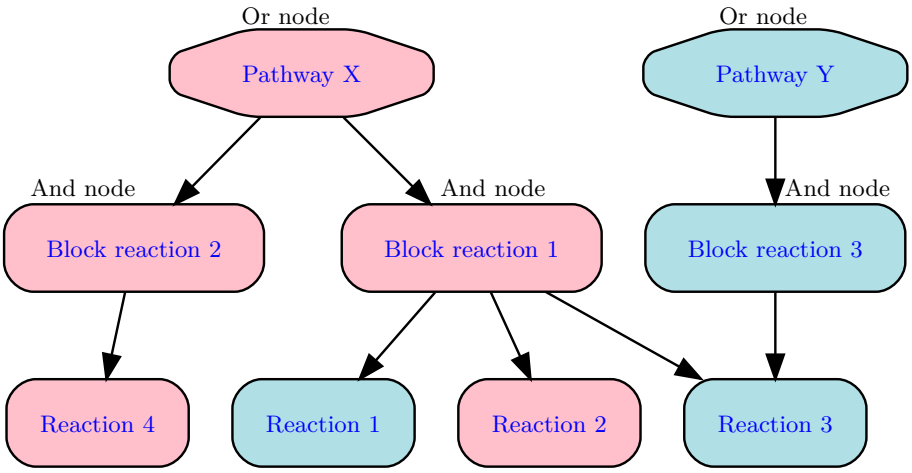

Fig. 2: Guided metabolic network reconstruction. Octagon shape is used for “or node ” and a rounded box for “and node ”. A red backgrounds means a knowledge is not present as opposite a cyan color means a present knowledge. Pathway X and Y are required on studied organism. Bio-annotator point of view the pathway X is certainly present as a specific reaction .

3 Benefits of applying an oriented-object expert system

Biological facts are stored into many bio-data warehouse, which are cross-referenced to reason over. For a computer scientist such information fit an oriented object approach. An oriented object programming ease in handling the data. Each kind of data corresponds to a class. With the use of an object relational mapping, objects are automatically pull to a typed information from a data warehouse. An object is an instance of the corresponding class. A computer scientist will apply a formula, an algorithm, a reasoning, depending of the current information type used.
We do not know in advance which type object the end used will put. To communicate between these different type of objects, the vocabulary is described by the use of interface. Interface provides a low coupling to have a well-structured computer system. Interface impose communication protocol, a partial set of attributes from classes are shared through any type of object. In other hand interface avoid multi-inheritance issues and provide a standardized vocabulary through any type of object. The use of inheritance Interface define common accessor and mutator following the lowerCamelCase convention syntax (i.e getX, setX . . .), attribute name are extratable in a standard way from various objects. We describe three generics type of facts: Knowledge, Prediction and Assertion (Figure 3) .

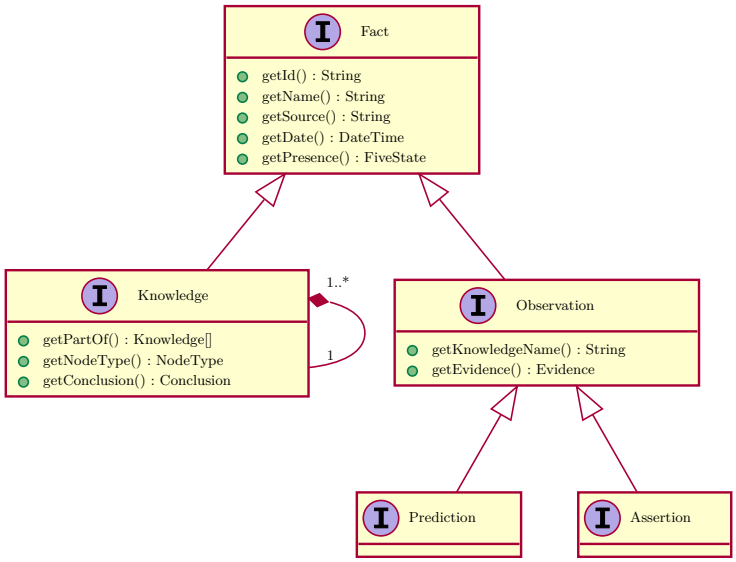

Fig. 3: Generic Fact model. Knowledge are used for a trusted information, Prediction for bioinformatics and bio-annotator observation and Assertion for an empiric observation. Predictions and Assertions results are facing to Knowledges.

3.1 Reasoning over structured data

Logic point of view a class is a structure. An instance of a class (an object) is a theory. An object can to contains others objects, that means a theory is a graph of smaller theories. Each theory are defined by a coherent set of values. A class C is 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 sub-theory can to be inferred to their upper theory following a Belnap logic (see Figure 1). In a such case a "AND" node have the following priority evaluation: FALSE > BOTH > NONE > TRUE. As example if one of sub-theory linked to a "AND" node is false the upper theory is false. Priority evaluation for "OR" node is: TRUE > NONE > BOTH > FALSE.

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	F	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

Point of view of a bio-annotator they are some cases not well cover by this logic. Indeed if a required pathway X has at least one reaction specific of the reaction block (not shared to another reactions blocks) then corresponding reaction block is present which infer the presence to the pathway X. That means we do not know yet what happen in this cases but the studied organisms has the required ability. By modifying "AND" node priority evaluation to TRUE > FALSE > BOTH > NONE in these cases we can to have an optimistic Belnap logic.

3.2 Hierarchical knowledge representation

Handling fact form a directed acyclic graph. A pathway contains one or more pathway variant. A pathway variant contains one or more reaction block. And reaction block contains one or more enzymatic reaction, these enzymatic reaction can to appear in one or more reaction block (as the reaction 3 in Figure 2). These knowledge need to be studied to know if is present or absent in the given organism. Prediction fact will infer in a bottom-up way while assertion fact will do it in a top-down way. Once all facts are spreads over the DAG of knowledge a refined conclusion is put on each knowledge. This conclusion will to oppose an assertion fact with a prediction fact comparing a four logic value to an another following the conclusion table 4.

Assertion / Presence	TRUE	FALSE	BOTH	UNKNOWN
REQUIRED	Confirmed P.	Unexpected A.	Contradictory A.	Missing
AVOIDED	Unexpected A.	Confirmed A.	Contradictory P.	Confirmed A.
BOTH	Ambiguous P.	Ambiguous A.	Ambiguous C.	Ambiguous
UNKNOWN	Unconfirmed P.	Unconfirmed A.	Unconfirmed C.	Unknown

Legend

A.	Absence
P.	Presence
C.	Contradictory

Fig. 4: Conclusion table.

4 Conclusion and future work

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. Belnap Jr, N.D.: A useful four-valued logic. In: Modern uses of multiple-valued logic, pp. 5–37. Springer (1977)
3. Bochner, B.R., Barnby, D.W.: Biolog. Website, http://www.biolog.com/products-static/phenotype_microbial_cells_overview.php
4. 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)
5. Mazumder, R., Natale, D.A., Julio, J.A.E., Yeh, L.S., Wu, C.H.: Community annotation in biology. Biol Direct 5, 12 (2010)
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 research p. gkr1023 (2011)
7. 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)