GROOLS: reactive graph reasoning for genome annotation through biological processes

Jonathan Mercier*, Adrien Josso , Claudine Médigue , and David Vallenet*

CEA, Genoscope, LABGeM, 2 rue Gaston Crémieux, 91057 Evry, France
Université d'Evry Val-d'Essonne, Evry, France
CNRS-UMR 8030, Evry, France
Université Paris-Saclay, Evry, France
Corresponding authors: jmercier@genoscope.cns.fr, vallenet@genoscope.cns.fr

Compiled March 17, 2017

Motivation: High quality functional annotation is essential for understanding the phenotypic consequences encoded in a genome. Despite improvements in bioinformatics methods, millions of sequences in databanks are not assigned reliable functions. The curation of protein functions in the context of biological processes is a way to improve and validate their annotation.
Results: We developed an expert system using paraconsistent logic, named GROOLS (Genomic Rule Object-Oriented Logic System), that assists biocurators in functional annotation of proteins. GROOLS evaluates the completeness and the consistency of predicted functions through biological processes like metabolic pathways. At the end of the reasoning, conclusions are assigned to biological process components and highlight confirmed/missing predictions and potential conflicts with expectations (e.g. growth phenotype experiments). Results on 14 microbial organisms are presented.
Availability: GROOLS is implemented as a standalone Java application. GROOLS is free open source software distributed under CeCILL license and can be downloaded from https://github.com/Grools.
Supplementary information: Detailed results are available at https://www.genoscope.cns.fr/agc/grools

http://dx.doi.org/XX.XXXXXX

1. INTRODUCTION

Assigning functions to all predicted proteins from genome sequencing projects remains challenging despite improvements in bioinformatics methods as illustrated by the second critical assessment of functional annotation initiative (CAFA2) [1]. Millions of protein entries are thus not assigned reliable functions in databanks due to the lack of trustworthy annotations and the drawbacks of homology-based predictions [2]. Using this genomic data, genome-scale metabolic networks can be obtained and contain all of the known metabolic reactions in a given organism and the genes that encode each enzyme. High quality functional annotation is thus an essential step to correctly predict the chemical transformations that may occur in a living cell. From these networks, metabolic models can be derived to study metabolic and growth functions notably by using flux balance analysis [3]. The predictive power of these models is very sensitive to incomplete knowledge and their manual curation is time-intensive even if several algorithms can be used to predict which reactions are missing by comparing *in silico* growth simulations to experimental results [4].

In order to automatically annotate protein sequences with a high degree of accuracy (i.e. by limiting annotation error propagation between homologous proteins), the UniProt consortium has developed the UniRule system [5]. It contains a set of annotation rules that are defined by biocurators using UniProtKB/Swiss-Prot annotations as a gold standard. These rules are based on the presence of specific protein signatures together with taxonomic constraint to predict biological features and functions of unreviewed proteins of the UniProtKB/TrEMBL section. The process annotates proteins independently without considering other predicted functions in the studied organism. Improving and validating functional annotation of proteins in a metabolic context was shown of interest in several microbial genome annotation systems such as SEED [6], MicroScope [7, 8] and IMG [9]. Similarly to UniRule that predicts protein functions, Genome Properties [10, 11] and IMG [12] systems have defined rules to predict biological processes. These rules are a mix of disjunctions and conjunctions of process components that are hierarchically organized to represent, for example, a pathway composed of reactions that are catalyzed by enzymes made of polypeptides in complex. Rules are evaluated using protein annotations as facts to determine if a minimal set of required functions is predicted to infer the presence or absence

of a process in an organism. In the IMG system, additional rules based on biological processes are used to predict phenotypes like auxotrophy/prototrophy for biosynthetic pathways or compound utilization for degradation pathways.

Here, we present an expert system, named GROOLS (Genomic Rule Object-Oriented Logic System), that uses paraconsistent logic to assist biocurators in functional annotation of genes. The GROOLS method is inspired from a previous work about the HERBS software (A. Viari *et al.*, unpublished data) that was presented in 2003 during a workshop of the HAMAP project [13]. The objective is to automate human expert reasoning applied in the curation of genome annotation. Similarly to previously mentioned systems (i.e. Genome Properties and IMG), GROOLS evaluates the completeness and the consistency of predicted functions through biological processes like metabolic pathways. Our objective was to propose a framework made of a generic data model to embrace various sources of biological process definitions. Moreover, the tool should consider potential contradictions or incomplete information and should make a clear distinction between what is predicted and what is expected during the reasoning. Expectations could be background knowledge on the organism, experimental results like growth phenotypes or simply biological hypotheses provided by users. At the end of the reasoning, conclusions should highlight confirmed/missing predictions and potential conflicts between predictions and expectations. Following a description of the GROOLS method, results on 14 microbial organisms, having available growth phenotype data, are presented and illustrated by two examples. Finally, we conclude on the GROOLS functionalities, potential applications and improvements.

2. METHODS

A. Biological knowledge representation

As illustrated in Figure 1, biological knowledge in GROOLS is represented as a conceptual graph [14] with two types of nodes:

- Prior Knowledge (PK) concepts represent biological processes and their components (e.g. a metabolic pathway and its reactions)
- Observation (O) concepts are information about PK concepts in a given organism.

The hierarchical organization of PK concepts is made by two types of edges in the conceptual graph:

- 'part' relations describe compositions (e.g. a metabolic pathway is composed of a set of reactions)
- 'subtype' relations describe generalizations/specializations (e.g. Pathways 1.1 and 1.2 are two variants of the generalized metabolic pathway 1).

Edges are directed and labelled with the suffix "-of" (i.e. 'part-of' and 'subtype-of') when the relation is from the parent concept to the child concept and, conversely, with the prefix "has-" (i.e. 'has-part' and 'has-subtype').

To give clues about the existence of PK concepts in a given organism, three types of Observation are used in GROOLS:

- Computations are bioinformatics predictions (e.g. genome annotation methods that predict genes encoding enzymes and their corresponding catalyzed reactions)
- Experimentations are empirical evidences (e.g. growth experiments with defined metabolites as sole carbon and energy source)
- Curations are human expertises supported by direct or indirect evidences (e.g. the validation of a bioinformatics prediction with additional literature evidences)

Finally, Observations are linked to PK concepts with two types of relations :

- 'prediction-of' edges for Computation and Curation observations
- 'expectation-of' edges for Experimentation and Curation observations.

This GROOLS conceptual graph model is generic enough to capture the overall background knowledge on biological processes represented by PK concepts (e.g. metabolic pathways from all domains of life) and organism specific knowledge represented by Observations that come from experimentations or its genome annotation. Each observation corresponds either to an assertion or a rejection of a PK concept. Such states are represented by the true (t) and false (f) values also named truth values in classical logic. Observations will be propagated through the conceptual graph to assign PK concepts prediction or expectation values (see the "Reactive reasoning" paragraph). These values answer the following questions: "Is the PK concept predicted in the studied organism?" and "Is it expected in the organism?".

B. Observation and Prior Knowledge truth value powersets

In biology, several experimentations or bioinformatics predictions may be related to a same PK concept. To capture facts from this observable space, Observations are grouped by their PK target and relation type (i.e. 'prediction-of' or 'expectation-of') resulting from an Observation truth value powerset of order 2 ($\mathbb{P}(2)$) composed of: {t} (true-only Observations), {f} (false-only Observations), {t,f} (true and false Observations) or $\{\emptyset\}$ (no Observation). These truth value sets of Observations (Otvs) are associated to the truth degrees 1, 0, $\frac{1}{2}$, 0 and to the falsehood degrees 0, 1, $\frac{1}{2}$, 0 for the {t}, {f}, {t,f} and $\{\emptyset\}$ sets, respectively.

In the conceptual graph, Observation sets are propagated directly to PK but also indirectly through PK concept relations. Thus, both prediction and expectation values of PK hold sets of direct and indirect Observations leading to a powerset of order 4, $\mathbb{P}(4)$ (i.e. 16 sets listed in Table 1). This paraconsistent logic uses an ensemblist representation of true and false value combination. It allows us to deal, without approximation, with uncertainty or contradiction for PK predictions and expectations. An algebraic structure behind these 16 subsets was proposed by Shramko *et al.* [15, 16] and is known as a trilattice SIXTEEN_3 with the following distinct partial orders: the truth, the falsehood and the degree of information. In GROOLS, the degree of truth/falsehood of each subset is computed by the sum of Otvs truth/false degrees present in the powerset divided by the degree of information. Finally, to determine the degree of belief, PK predictions and expectations are ranked by decreasing truth degree, then by increasing degree of falsehood and information, in case of equality: $\{\{t\}\}$, $\{\{t\},\{t,f\}\}$, $\{\{\emptyset\},\{t\}\}$, $\{\{\emptyset\},\{t\},\{t,f\}\}$, $\{\{t,f\}\}$, $\{\{t\},\{f\}\}$, $\{\{t\},\{f\},\{t,f\}\}$, $\{\{\emptyset\},\{t\},\{f\},\{t,f\}\}$, $\{\{\emptyset\},\{t\},\{f\}\}$, $\{\{\emptyset\},\{t,f\}\}$, $\{\{f\},\{t,f\}\}$, $\{\{\emptyset\},\{f\},\{t,f\}\}$, $\{\emptyset\}$, $\{\{\emptyset\}\}$, $\{\{\emptyset\},\{f\}\}$, $\{\{f\}\}$ (Table 1).

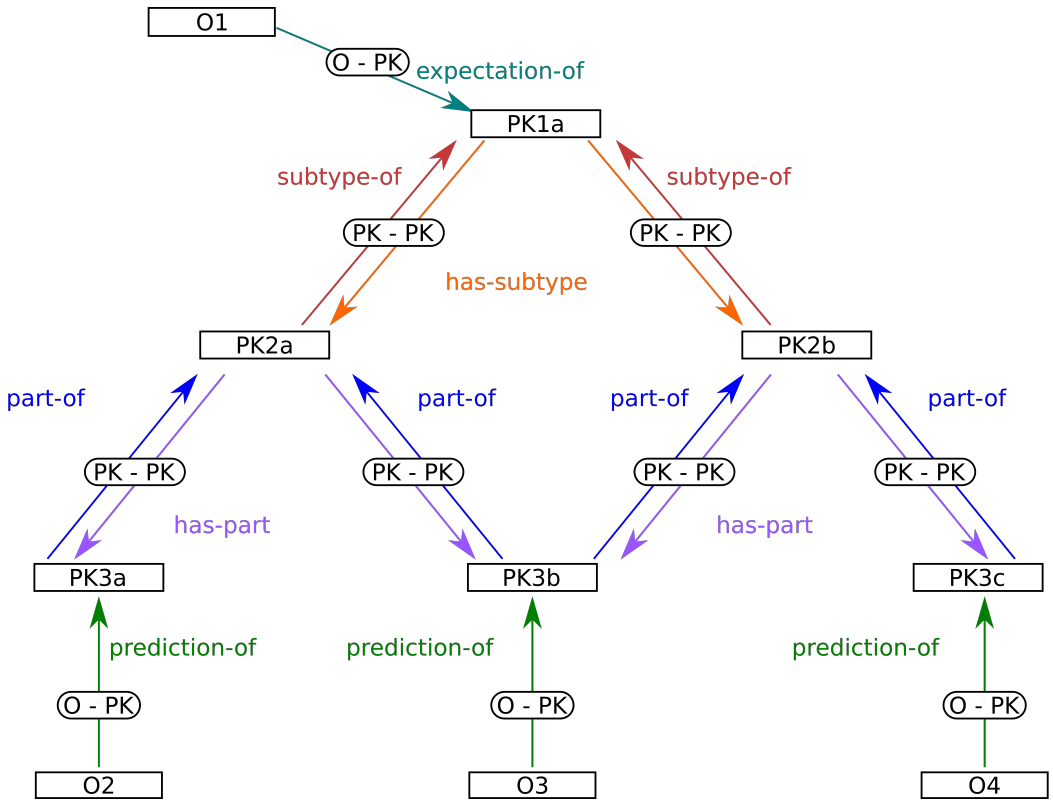

Fig. 1. GROOLS conceptual graph model. Boxes correspond to Prior Knowledge (PK) and Observation (O) concept nodes. Directed edges between PK concepts or between O and PK concepts are labelled PK-PK and O-PK, respectively. The name of the different types of edges are written next to the arrows.

Table 1. The sixteen truth value sets and their attributes

Attributes / Truth value sets	degree of				approximation	
	truth	falsehood	information	belief rank	prediction	expectation
$\emptyset$	0	0	0	13	NONE	NONE
$\{\{\emptyset\}\}$	0	0	1	14	NONE	NONE
$\{\{f\}\}$	0	$\frac{1}{1}$	1	16	FALSE	FALSE
$\{\{t\}\}$	$\frac{1}{1}$	0	1	1	TRUE	TRUE
$\{\{t,f\}\}$	$\frac{\frac{1}{2}}{1}$	$\frac{\frac{1}{2}}{1}$	1	5	BOTH	BOTH
$\{\{\emptyset\},\{f\}\}$	0	$\frac{1}{2}$	2	15	NONE	FALSE
$\{\{\emptyset\},\{t\}\}$	$\frac{1}{2}$	0	2	3	NONE	TRUE
$\{\{\emptyset\},\{t,f\}\}$	$\frac{\frac{1}{2}}{2}$	$\frac{\frac{1}{2}}{2}$	2	10	BOTH	BOTH
$\{\{f\},\{t\}\}$	$\frac{1}{2}$	$\frac{1}{2}$	2	6	BOTH	BOTH
$\{\{f\},\{t,f\}\}$	$\frac{\frac{1}{2}}{2}$	$\frac{1+\frac{1}{2}}{2}$	2	11	BOTH	BOTH
$\{\{t\},\{t,f\}\}$	$\frac{1+\frac{1}{2}}{2}$	$\frac{\frac{1}{2}}{2}$	2	2	BOTH	BOTH
$\{\{f\},\{t\},\{t,f\}\}$	$\frac{1+\frac{1}{2}}{3}$	$\frac{1+\frac{1}{2}}{3}$	3	7	BOTH	BOTH
$\{\{\emptyset\},\{f\},\{t\}\}$	$\frac{1}{3}$	$\frac{1}{3}$	3	9	BOTH	BOTH
$\{\{\emptyset\},\{f\},\{t,f\}\}$	$\frac{\frac{1}{2}}{3}$	$\frac{1+\frac{1}{2}}{3}$	3	12	BOTH	BOTH
$\{\{\emptyset\},\{t\},\{t,f\}\}$	$\frac{1+\frac{1}{2}}{3}$	$\frac{\frac{1}{2}}{3}$	3	4	BOTH	BOTH
$\{\{\emptyset\},\{t\},\{f\},\{t,f\}\}$	$\frac{1+\frac{1}{2}}{4}$	$\frac{1+\frac{1}{2}}{4}$	4	8	BOTH	BOTH

C. Reactive reasoning

The GROOLS reasoner propagates information from Observation to PK concepts and between PK using inference rules in three steps : propagation of predictions, propagation of expectations and PK conclusions. Thanks to the reactive nature of GROOLS reasoning, declaration of new Observations can be made at any time and logical consequences are propagated locally in the graph without re-evaluating all PK predictions and expectations.

C.1. Propagation of predictions

PK prediction values are assigned by collecting truth value sets from direct Observations (Computation or Curation type) and child PK predictions. The algorithm propagates predictions from the leaves to the roots of the conceptual graph using two distinct rules: one for 'has-part' relations (Rule 1) and another for 'has-subtype' relations (Rule 2). For the Rule 1, PK predictions are the union of Otvs (Observation truth value sets) and Ptvs (Prediction truth value sets) from child PK concepts. In Rule 2, PK predictions are the union of Otvs and Ptvs from child PK concepts having the greatest degree of belief (i.e. the smallest belief rank).

Rule 1. Prediction propagation with 'part-of' relation

Input: parent is a PriorKnowledge
Output: prediction set of parent
1: Otvs: { tvs | obs prediction-of parent ∧ tvs: obs.truthvalue }
▷ Otvs is a truth value set from $\mathbb{P}(2)$
2: children: { pk | pk part-of parent }
3: Ptvs : { tvs | pk ∈ children ∧ tvs ∈ pk.prediction) }
▷ Ptvs is a truth value set from $\mathbb{P}(4)$

$$parent.prediction \begin{cases} \{Otvs\} \cup Ptvs, & \text{if } Otvs \neq \emptyset \wedge Ptvs \neq \emptyset \\ Ptvs, & \text{if } Ptvs \neq \emptyset \\ \{Otvs\}, & \text{if } Otvs \neq \emptyset \\ \{\emptyset\}, & \text{otherwise} \end{cases}$$

Rule 2. Prediction propagation with 'subtype-of' relation

Input: parent is a PriorKnowledge
Output: prediction set of parent
1: Otvs: { tvs | obs prediction-of parent ∧ tvs: obs.truthvalue }
2: children: { pk | pk subtype-of parent }
3: Ptvs : { tvs | pk ∈ children ∧ tvs: pk.prediction) }

$$parent.prediction \begin{cases} \{Otvs\} \cup f_{maxBelief}(Ptvs), & \text{if Otvs} \neq \emptyset \\ & \wedge \text{ Ptvs} \neq \emptyset \\ f_{maxBelief}(Ptvs), & \text{if } Ptvs \neq \emptyset \\ \{Otvs\}, & \text{if } Otvs \neq \emptyset \\ \{\emptyset\}, & \text{otherwise} \end{cases}$$

▷ maxBelief function return the truth value set from $\mathbb{P}(4)$ having the smallest rank of belief

C.2. Propagation of expectations

In a second step, PK expectations are propagated from the roots to the leaves of the graph by collecting truth value sets from direct or indirect Observations (Experimentation or Curation type). For the 'has-part' relations, PK expectations are the union of Otvs and Etvs (Expectation truth value sets) from parent PK concepts (Rule 3). For the 'has-subtype' relations, parent PK expectations are propagated to child PK that have the greatest degree of belief for their prediction with the exception of false-only expectations ({{f}}) that are propagated to all child PK (Rule 4).

Rule 3. Expectation propagation with 'has-part' relation

Input: child is a PriorKnowledge
Output: expectation set of child
1: Otvs: { tvs | obs expectation-of child ∧ tvs: obs.truthvalue }
2: parents: { pk | pk has-part child }
3: Etvs : { tvs | pk ∈ parents ∧ tvs ∈ pk.expectation) }
▷ Etvs is a truth value set from $\mathbb{P}(4)$

$$child.expectation \begin{cases} \{Otvs\} \cup Etvs, & \text{if } Otvs \neq \emptyset \wedge Etvs \neq \emptyset \\ Etvs, & \text{if } Etvs \neq \emptyset \\ \{Otvs\}, & \text{if } Otvs \neq \emptyset \\ \{\emptyset\}, & \text{otherwise} \end{cases}$$

Rule 4. Expectation propagation with 'has-subtype' relation

Input: child is a PriorKnowledge
Output: expectation set of child
1: Otvs: { tvs | obs expectation-of child ∧ tvs: obs.truthvalue }
2: parents: { pk | pk has-subtype child }
3: Etvs: { tvs | pk ∈ parents ∧ tvs: pk.expectation) }
4: children: { pk | pk ∈ parents ∧ pk subtype-of parent }
5: Ptvs: { tvs | pk ∈ children ∧ tvs: pk.prediction) }
6: PmaxB: $f_{maxBelief}$(Ptvs)

$$child.expectation \begin{cases} \{\{f\}\}, & \text{if } Etvs = \{\{f\}\} \\ \{Otvs\} \cup Etvs, & \text{if Otvs} \neq \emptyset \\ & \wedge \text{ Etvs} \neq \emptyset \\ & \wedge \text{ child.prediction} = \text{PmaxB} \\ Etvs, & \text{if Etvs} \neq \emptyset \\ & \wedge \text{ child.prediction} = \text{PmaxB} \\ \{Otvs\}, & \text{if } Otvs \neq \emptyset \\ \{\emptyset\}, & \text{otherwise} \end{cases}$$

C.3. PK conclusions

At the end of the reasoning, to ease human interpretation of PK predictions and expectations, truth value sets are approximated in four values: TRUE, FALSE, BOTH, NONE (Table 1). When both true and false values are present, the approximated value is BOTH to express contradiction. For predictions, if the set contains an unknown value ({∅}) and a true or false value, the approximated value is NONE meaning that some parts of the concept are unknown. For expectations, unknown values are not considered in the approximation as they do not represent missing part of a concept but lack of information about a higher concept among others having empirical evidences (true or false expectation values). Thus {{∅}{t}} and {{∅}{f}} expectation sets are approximated to TRUE and FALSE, respectively.

The combination of these approximated values of PK predictions and expectations lead to 16 different states (Table 2). They express a conclusion about the presence/absence of a concept in an organism according to experiments and genome annotations and point out potential contradictions and uncertainties. For examples, a "Confirmed Presence" is used when the concept is both predicted and expected. An "Unexpected Presence" corresponds to a predicted process for which direct or indirect experimental evidences indicate that it does not occur in the studied organism. If experiments contradict themselves (BOTH expectation) while the process is predicted, the conclusion is an "Ambiguous Presence". Finally, as a last example, the "Missing" state is an absence of prediction (NONE prediction) while the process is experimentally observed. In this case, the genome annotation should be improved to find candidate proteins for the missing functions.

Table 2. Conclusion table

Expectation \ Prediction	TRUE	FALSE	BOTH	NONE
TRUE	Confirmed P.	Unexpected A.	Contradictory A.	Missing
FALSE	Unexpected P.	Confirmed A.	Contradictory P.	Absent
BOTH	Ambiguous P.	Ambiguous A.	Ambiguous C.	Ambiguous
NONE	Unconfirmed P.	Unconfirmed A.	Unconfirmed C.	Unexplained

P.: Presence, A.: Absence, C.: Contradiction

C.4. Dispensable, falsehood and specific modes

Three additional reasoning modes have been developed to handle particular cases in the reasoning. In the Dispensable mode, PK concepts can be flagged as non-essential meaning that their presence is not required to realize a biological process. In this case, predictions and expectations are not propagated to PK concepts labelled as "Dispensable". In the Falsehood mode, leaf PK concepts (i.e. concepts without 'has-part' and 'has-subtype' relations) with no direct Observations have their prediction assigned to false-only ({{f}}). This reasoning mode assumes that all leaf concepts of the graph are predictable, and thus, an absence of Observation corresponds to a false prediction. In the Specific mode, a PK is labelled as "Specific" when it is linked to only one parent PK by a 'part-of' relation. Moreover, in this mode, Rule 1 (propagation of child PK predictions to parent PK using 'part' relations) is adapted: child PK with unknown prediction ({{∅}}) are not considered when at least one other child PK has a true-only prediction ({{t}}). Using this Specific mode, reasoning is less strict as partial evidences on child PK are sufficient to predict a parent PK.

D. Biological data

Prior Knowledge (PK) concepts represent biological processes and their components (e.g. a metabolic pathway and its reactions). Genome Properties [10] and Unipathway [17] resources are currently supported by GROOLS to reconstruct two distinct conceptual graphs.

From the Genome Properties description file (version 3.2, GenProp_3.2_release.RDF), "Properties", "Components" and "Evidences" were extracted and transformed in PK concepts. 'Part' relations were established between "Properties" and "Components", and 'subtype' relations between "Components" and "Evidences". In the Genome Properties model, "Evidences" correspond either to pathway variants or predictors (i.e. leaf concepts in the graph). Predictors are protein family Hidden Markov Models (HMMs) from PFAM [18] and TIGRFAM [11] databases. Several leaf "Evidences" linked to a "Component" represent distinct protein families performing the same function (e.g. isoenzymes).

From the UniPathway Open Biological Ontology (OBO) file, UPA (pathways), ULS (linear sub-pathways), UER (enzymatic reactions) and UCR (chemical reactions) were extracted and transformed in PK concepts. For ULS linked to a same UPA, additional concepts, named Variants, were added when several paths of ULS, that share input/output compounds, exist. Other Variants were made for UER that are specializations of generic UER (i.e. "has_alternate_enzymatic_reaction" relations in the OBO file). According to "is_a" and "super_pathway" relations in the OBO file, UPA were linked together by 'subtype' relations in the conceptual graph. 'Subtype' relations were also used to link Variants to UPA or UER. Other relations between PK concepts were converted to 'part' relations.

To populate Genome Properties conceptual graph, Observations of Computation type can be extracted from UniProtKB database [5] in which each protein entry is linked to predicted Pfam and TIGRFAM families. For UniPathway, UniProtKB annotations can be used and inserted in the conceptual graph as Observations of UER using their identifiers (when provided in the protein entry) or their cross-referenced EC (Enzyme Commision) numbers. Annotations from the MicroScope platform [8] can be used as a second resource of UniPathway Observations. MicroScope contains automatic and curated annotations of microbial genomes. Similarly to UniProtKB, protein annotated with EC numbers are converted in UER Observations. In addition, Observations on UCR are made using cross-referenced MetaCyc [19] or RHEA [20] reactions that are annotated in MicroScope protein entries. Automatic annotations from MicroScope are considered as Observations of Computation type, whereas curated annotations are considered as Observations of Curation type (i.e. Observation truth value sets are propagated both on PK predictions and expectations). A specific case concerns MicroScope pseudogenes which predicted original functions are considered as false Observations of UCRs and UERs (i.e. corresponding reactions should not occur anymore in the organism as pseudogenes do not produce functional proteins).

GROOLS can integrate Phenotype MicroArrays (PMs) BIOLOG experiments as Observations of Experimentation type. This quantitative growth phenotype data should be first discretized using the omp R package [21] with "grofit" aggregation method and weak discretization (-a, -w and -z options of run_opm.R program). Using Observations with a true (growth) or false (no growth) truth value, the results obtained with carbon and nitrogen sources (PM1, 2 and 3 plates) are then linked to their corresponding Genome Properties (16 pathways) or Unipathway (37 pathways) PK concepts (i.e. the metabolic pathway for the degradation of the tested compound). For ambiguous discretizations (NA values), both true and false Observation concepts are created. Regarding prototrophic organisms, true Observations on amino acid biosynthesis pathways can be added as a second source of Experimentation corresponding to 15 pathways in Genome Properties and 20 in UniPathway.

E. Statistics

Statistics were computed to evaluate predictions in confrontation to biological expectations. In this evaluation, only concepts with univocal expectations ({{t}} and {{f}} values) are considered. Concepts are classified using their conclusion value as follow:

- True Positive (TP): 'Confirmed Presence'
- True Negative (TN): 'Confirmed Absence' and 'Absent'
- False Positive (FP): 'Unexpected Presence' and 'Contradictory Presence'

- False Negative (FN): 'Missing', 'Unexpected Absence' and 'Contradictory Absence' concepts.

Accuracy (ACC) is computed as follow:

$$ACC = \frac{(TP + TN)}{(TP + FP + FN + TN)}$$

3. IMPLEMENTATION

GROOLS is built on object-oriented logic programming. This paradigm facilitates the representation of biological knowledge with Prior Knowledge concepts and their relations. A first prototype of GROOLS (https://github.com/Grools/grools-drools-checker) was written using the DROOLS framework (https://www.drools.org, a Business Rules Management System based on an object-oriented rule language). In the present version of GROOLS software (https://github.com/Grools/grools-application), rules were directly implemented in Java with functional programming to tackle performance issues during the graph reasoning. The reasoner is available in a separate library (https://github.com/Grools/grools-reasoner).

GROOLS contains plugins to construct the PK graph from two databases: UniPathway (in OBO format) and Genome Properties (in non-standard format). Using this plugin system, database file content can be improved by users. Moreover, other resources on biological processes can be integrated without changing the reasoning library.

For Observations, an unified CSV file format is used to represent Computation, Curation and Experimentation data. Each row is composed of seven fields to describe an Observation: "Name", "EvidenceFor" (name of the related Prior Knowledge), "Type" (Computation, Curation or Experimentation), "isPresent" (true or false), "Source" (e.g. UniProtKB, MicroScope, BIOLOG), "Label" and "Description".

Three shell scripts are provided as example to run GROOLS on Genome Properties using UniProtKB proteomes ("uniprot_genpropToGrools.sh" script) and on UniPathway using MicroScope or UniProtKB annotations ("microscope_upaToGrools.sh" and "uniprot_upaToGrools.sh", respectively).

GROOLS results are provided in CSV format and in HTML pages containing graphical representations of the conceptual graph for each biological process. GROOLS execution time is about 1 to 2 minutes to completely analyze a genome. The reasoning part takes only few seconds while most of the execution time is spent in data extraction and graphical output generation.

4. RESULTS

A. GROOLS reasoning test case

As a test case, GROOLS was executed on 14 organisms for which UniProtKB and MicroScope annotations are available (see Supplementary Data 1). The reasoning was made on Genome Properties and UniPathway conceptual graphs, enabling or not the falsehood (for Genome Properties) and specific (for UniPathway) modes. In addition, the dispensable mode was enabled for Genome Properties as this resource contains information about non-essential components of biological processes. For Observations of Experimentation type, growth phenotypes (PM BIOLOG technology) were retrieved for these organisms on the Microme project website (http://www.microme.eu) and linked to their corresponding degradation pathway PK concepts. Given that studied organisms are prototroph, additional Observations with true truth value were added on amino acid biosynthesis pathway concepts (i.e. these pathways are expected in the organisms). Detailed results on the 14 studied organisms are available at this Web address: https://www.genoscope.cns.fr/agc/grools. Statistics on GROOLS conclusions are reported in Table 3. Accuracies of pathway and functional unit predictions were computed only for pathways with univocal expectations (i.e. pathway concepts with true-only or false-only Observations of Experimentation type).

For Genome Properties, the evaluation of functional unit predictions in confrontation to pathway expectations (i.e. from BIOLOG data and amino acid biosynthesis pathways) gives high level of accuracy (79.14%), reflecting the exquisitely sensitive searches of TIGRFAM models. In the falsehood mode, pathway prediction accuracy decreases from 78.80% to 67.93%. As expected, we observe that contradictions on pathway predictions (i.e. conclusions with BOTH predictions) are only detected using the falsehood mode. In this mode, the prediction value of a leaf component (i.e. a functional unit) is set to false-only when any corresponding PFAM or TIGRFAM domain is detected among all organism proteins. Through the propagation of these false values, pathway predictions change then from NONE to FALSE or BOTH values. One advantage of the falsehood mode is thus to raise contradictions between predictions and expectations at the pathway level (i.e. 'Contradictory Absence' and 'Contradictory Presence' conclusions). Furthermore, using falsehood mode, absence of pathways and prediction contradictions can be detected even without expectations (i.e. 'Unconfirmed Absence' and 'Unconfirmed Contradiction' conclusions).

Using UniPathway as a source of biological processes, prediction accuracies for pathways and functional units are higher for MicroScope than for UniProtKB: about twice as many functional units are correctly predicted using MicroScope annotations. Indeed, MicroScope contains further annotations on enzymes. In addition to EC numbers, reactions identifiers from MetaCyc or RHEA are generally provided in MicroScope, which facilitate the mapping of Observations on UniPathway chemical reactions (UCRs). Moreover, the UniRule annotation system is still limited to few UniProtKB entries and many entries contain original submitted annotations that are very heterogeneous in quality. Beside, the specific mode shows a positive impact on prediction accuracy. For MicroScope annotations, pathway and functional unit accuracy values increase from 69.35% to 73.70% and from 82.22% to 84.15%, respectively. Indeed, the number of 'Missing' conclusions (i.e. false negative predictions) is reduced to half using the specific mode. In this mode, a pathway variant is assigned a true prediction if at least one specific reaction is predicted (i.e. it does not require all reactions to be predicted). Thus, the specific mode helps GROOLS reasoning to select the correct pathway variant (i.e. the one with the greatest degree of belief) for which pathway expectations are propagated.

For BIOLOG results, it should be noticed that 23% and 37% of Genome Properties and UniPathway pathways have contradictory expectations due to ambiguous discretizations or discordant experimental replicates. By taking into account contradictory experimental results, GROOLS may help biologists to deal with these inconsistencies by pointing out, for example, pathways with 'Ambiguous Presence' conclusion, which suggests that, since the pathway is predicted, the organism should be able to degrade the tested compound.

Detailed statistics on prediction accuracy for the 14 studied organisms are represented in Figure 2 (see Supplementary Data 1 for detailed counts). Results were obtained using Genome

Table 3. GROOLS statistics on pathways and functional units for Genome Properties and UniPathway. Statistics are computed for the 14 studied organisms.

		Genome Properties		UniPathway			
		UniProtKB	UniProtKB falsehood	UniProtKB	MicroScope	UniProtKB specific	MicroScope specific
Pathway conclusions	**Exp. / Pred.**						
Confirmed Presence	TRUE / TRUE	195	195	241	277	339	393
Unexpected Absence	TRUE / FALSE	0	12	0	0	0	0
Contradictory Absence	TRUE / BOTH	0	56	0	5	0	5
Missing	TRUE / NONE	68	0	232	191	137	80
Unexpected Presence	FALSE / TRUE	13	13	33	50	79	121
Confirmed Absence	FALSE / FALSE	0	66	0	0	0	0
Contradictory Presence	FALSE / BOTH	0	40	0	0	0	0
Absent	FALSE / NONE	106	0	275	258	229	187
Ambiguous Presence	BOTH / TRUE	15	15	47	68	121	178
Ambiguous Absence	BOTH / FALSE	0	13	0	0	0	0
Ambiguous Contradiction	BOTH / BOTH	0	24	0	3	0	3
Ambiguous	BOTH / NONE	37	0	243	218	172	120
Unconfirmed Presence	NONE / TRUE	1222	1222	728	946	1130	1409
Unconfirmed Absence	NONE / FALSE	0	3453	0	0	0	0
Unconfirmed Contradiction	NONE / BOTH	0	726	0	5	0	4
Unexplained	NONE / NONE	4238	59	2359	2137	1951	1658
Pathway statistics							
True Positive		184	184	200	238	272	310
True Negative		106	66	188	176	159	130
False Positive		13	53	19	31	48	77
False Negative		65	65	190	152	118	80
Accuracy		78.80%	67.93%	64.99%	69.35%	72.19%	73.70%
Functional unit statistics							
True Positive		1236	1236	969	1719	982	1742
True Negative		710	710	568	505	568	504
False Positive		117	117	100	139	100	140
False Negative		396	396	511	342	465	283
Accuracy		79.14%	79.14%	71.55%	82.22%	73.29%	84.15%

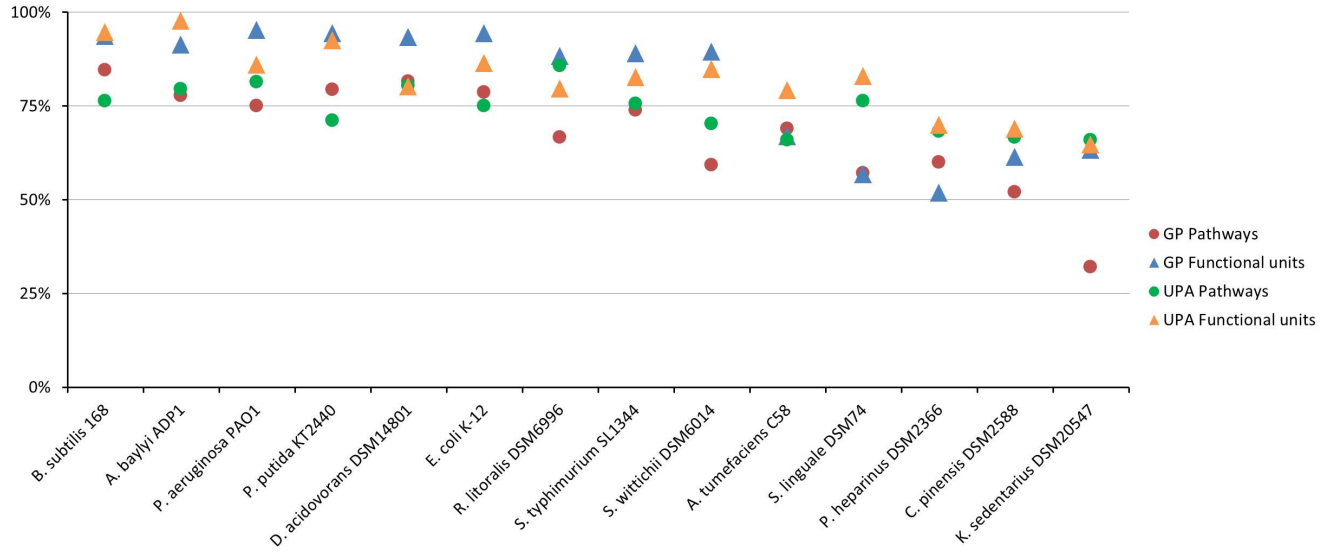

Fig. 2. Prediction accuracy for pathways and functional units from Genome Properties (GP) and UniPathway (UPA) in 14 organisms. For Genome Properties, GROOLS falsehood mode is activated. For UniPathway, the specific mode is activated and annotations are from MicroScope.

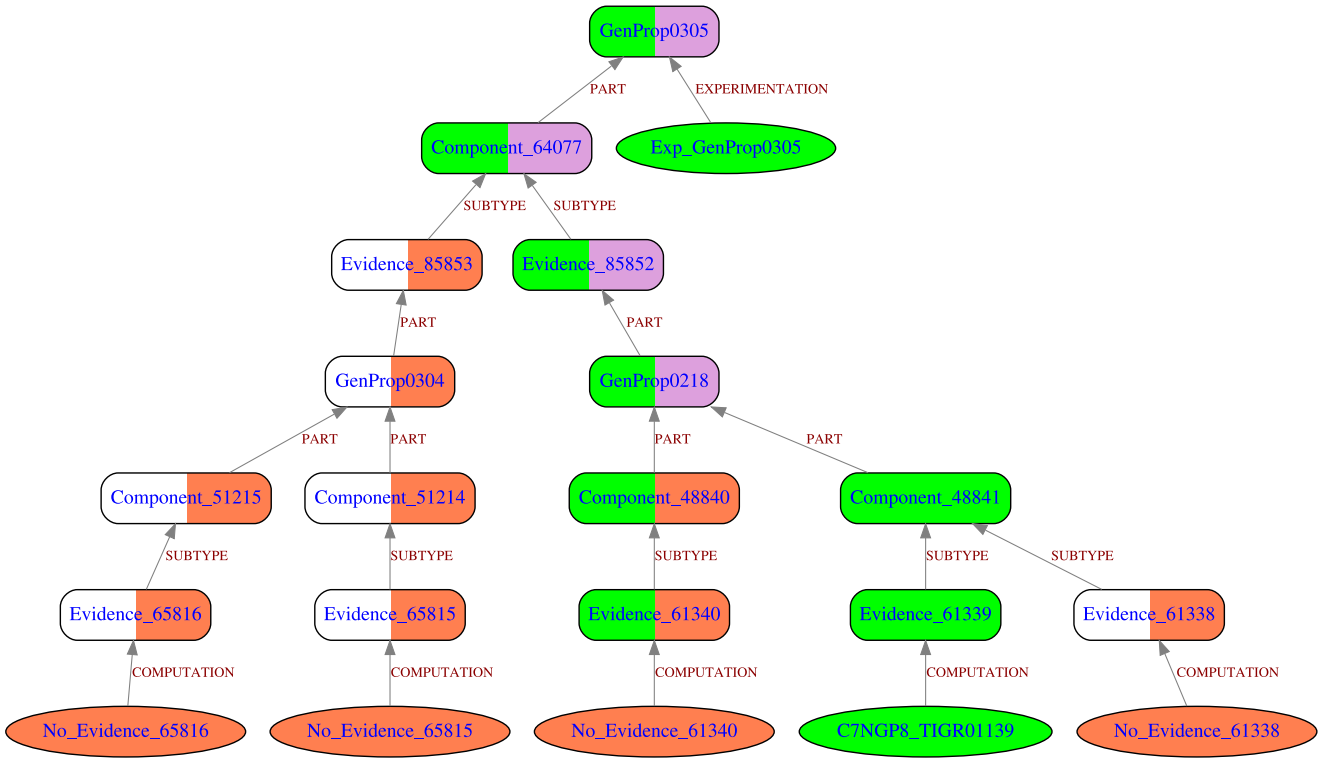

Fig. 3. GROOLS results for Genome Properties cysteine biosynthesis in *Kytococcus sedentarius* (falsehood mode reasoning). Rounded boxes are Prior Knowledge concepts and ovals are Observations. Edges between Observations and Prior Knowledge concepts are labelled by the type of Observation whereas, between Prior Knowledge concepts, labels correspond to relation types. The color code corresponds to TRUE (green), FALSE (red), BOTH (purple) and NONE (white) values. For Prior Knowledge concepts, the colored left part corresponds to expectation value and the right part to prediction value.

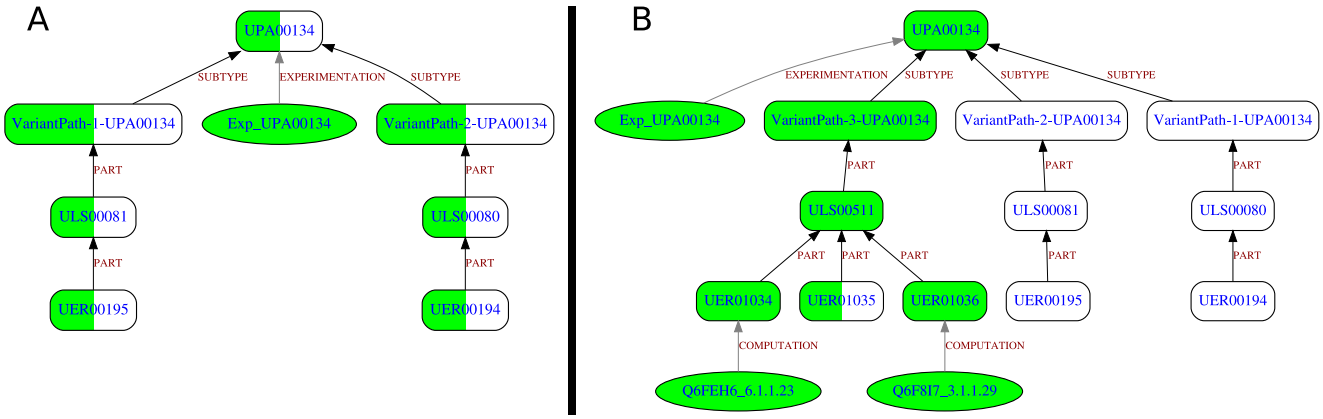

Fig. 4. GROOLS results for UniPathway asparagine biosynthesis in *Acinetobacter baylyi* (specific mode reasoning with UniProtKB annotations). Part A corresponds to the original pathway definition in UniPathway. Part B corresponds to the enhanced pathway definition with an additional variant. See Figure 3 for the legend.

Properties in falsehood mode and using UniPathway in specific mode with MicroScope annotations. As expected, model bacteria (*Escherichia coli* and *Bacillus subtilis*) or other well-studied organisms (*Acinetobacter baylyi*, *Pseudomonas putida* and *P. aeruginosa*) have high prediction accuracies for functional units (accuracy >85%) and pathways (accuracy >70%). After manual inspection of some results, many discrepancies were found to be annotation mapping problems with UniPathway reactions (i.e. EC numbers or MetaCyc/RHEA reactions are not correctly cross-referenced) or incomplete pathway definition (i.e. missing variant) as illustrated in the second example below. Other organisms, like *Chitinophaga pinensis* and *Kytococcus sedentarius*, have low prediction accuracies (<70%). These organisms are from phylum (i.e. Bacteroidetes and Actinobacteria) with limited metabolic knowledge. Potential new pathways and enzymes remain to be discovered or, more easily, predictors like TIGRFAM models could be improved to gain sensitivity as illustrated in the first example just below.

B. Genome Properties falsehood mode example

As depicted in Figure 3, cysteine biosynthesis pathway (GenProp0305) in *Kytococcus sedentarius* is flagged by GROOLS as 'Contradictory Absence' conclusion meaning that, although the pathway is expected in this organism, some predictions are in contradiction. Two pathway variants for cysteine biosynthesis are described in the Genome Properties model: GenProp0304 (cysteine biosynthesis, tRNA-dependent) and GenProp0218 (cysteine biosynthesis from serine). Among all functional units, only one is predicted in *K. sedentarius* (Evidence_61339 for TIGR01139 HMM) others are absent and have their prediction value set to false (falsehood mode). Rule 1 and Rule 2 of GROOLS reasoning were then applied to propagate predictions to parent concepts leading to 'false-only' and 'true and false' predictions for GenProp0304 and GenProp0218, respectively. Using rules 3 and 4, the Observation Exp_GenProp0305 (i.e. cysteine biosynthesis is expected in the organism) was propagated only through GenProp0218 as it has the greatest degree of belief for its prediction (i.e. GenProp0218 prediction value is {{t,f}} whereas GenProp0304 is {{f}}). This reasoning decision confirmed the presence of Evidence_61339 (TIGR01139) whereas Evidence_61340 (TIGR01172) is flagged as 'Unexpected Absence': any protein in *K. sedentarius* has significant alignment with TIGR01172 HMM. Such conclusion suggests that this missing function should be present in *K. sedentarius*. A quick investigation of TIGRFAM results showed that Ksed_26170 protein (C7NGP7 UniProtKB entry) shares an alignment with TIGR01172 HMM with a score of 200.3 that is just below the trusted cutoff (210.65). Thus, this protein is certainly a good candidate for the missing function and the TIGRFAM HMM should be updated accordingly to gain sensitivity in the detection of proteins from less-studied organisms like the actinobacteria *K. sedentarius*.

C. UniPathway specific mode example

In UniPathway, two variants made of one enzymatic reaction are described for the biosynthesis of asparagine (UPA00134). Using *Acinetobacer baylyi* UniProtKB annotations (Figure 4 part A), any observation is linked to the two enzymatic reactions (UER00194 and UER00195) (i.e. there is any protein annotated with the corresponding UERs). Since this pathway is expected in *A. baylyi* (Exp_UPA00134 Observation), GROOLS reasoning concludes that both variants and reactions could be missing (i.e. GROOLS is not able to choose between those two). Unfortunately, UniPathway is wrong because a third pathway variant is known and reported in other metabolic databases like MetaCyc [19]. It corresponds to a tRNA-dependent transamidation mechanism for the conversion of aspartate to asparagine. This variant made of 3 reactions was integrated in the UniPathway model and the GROOLS reasoning was relaunched (Figure 4 part B). Using this modified model, GROOLS correctly selects the third variant by assigning a 'Confirmed Presence' conclusion and detects a missing reaction (UER01035). This latter is catalyzed in *A. baylyi* by an heterotrimer made of GatA, B, C proteins which cannot be assigned to the UER01035 as the corresponding EC number (6.3.5.6) is missing in their UniProtKB annotation.

5. DISCUSSION

Following the initiative of Genome Properties [10] and IMG [9] systems, we present here a methodological framework for the evaluation of protein functions in the context of biological processes like metabolic pathways. One original aspect of our work is the use of paraconsistent logic to deal with uncertainty and contradiction in a discriminating way for both predictions and expectations. Biological processes are modeled in a generic and hierarchical representation of knowledge, which is based on a conceptual graph with 'part' and 'subtype' relations. The reasoning is made of four rules that are activated dynamically by the propagation of observations over the graph. A the end, we propose 16 different conclusion values that are assigned to all biological process components. GROOLS open source software implements this method and was tested over 14 organisms with growth phenotype data. Analyses were made for two biological process resources (Genome Properties and UniPathway) and two sources of annotation (UniProtKB and MicroScope). GROOLS results can be used to evaluate the overall accuracy of functional unit and pathway predictions according to organism experimental data. It allows users to compare different sources of annotations for a genome. Furthermore, using conclusion values, biocurators can quickly focused on missing or contradictory predictions to improve protein functional annotation. With the aim of building a metabolic model of an organism, and before applying time-intensive gap filling methods, GROOLS could thus be used to improve the initial metabolic network reconstruction in regards to the genome annotation and growth phenotype data.

To increase the diversity of biological processes in GROOLS, we plan to integrate data from MetaCyc [19] that contains a comprehensive list of experimentally elucidated metabolic pathways with variant definition. For the methodological part, a new type of relations, named 'avoid', could be implemented in GROOLS model and reasoning. Unlike other relations between PK, 'avoid' relation will indicate concept components that should not be present. This type of relation could be useful to model more complex phenotypes like, for example, 'strict anaerobe' for which anaerobic metabolic pathways should be predicted, but not aerobic ones. To conclude, GROOLS, by its reactive nature and its fast execution time, could be integrated in an annotation system (such as MicroScope) as a live annotation companion to guide biologists during the curation process of their genome: each time a protein annotation is updated, a new observation should be created in the GROOLS working memory and, then, would be propagated locally through the graph to alert users about impacts on biological process conclusions.

ACKNOWLEDGEMENTS

We would like to thank Anne Morgat for her valuable advice about UniPathway data, Thierry Lombardot for his help in SPARQL UniProtKB queries and Alain Viari for his support and constructive comments.

FUNDING

This work was not supported by any funding.

Conflict of Interest: none declared.

REFERENCES

1. Y. Jiang, T. R. Oron, W. T. Clark, A. R. Bankapur, D. DAndrea, R. Lepore, C. S. Funk, I. Kahanda, K. M. Verspoor, A. Ben-Hur *et al.*, "An expanded evaluation of protein function prediction methods shows an improvement in accuracy," Genome biology **17**, 184 (2016).

2. A. M. Schnoes, S. D. Brown, I. Dodevski, and P. C. Babbitt, "Annotation error in public databases: misannotation of molecular function in enzyme superfamilies," PLoS Comput Biol **5**, e1000605 (2009).

3. J. D. Orth, I. Thiele, and B. Ø. Palsson, "What is flux balance analysis?" Nature biotechnology **28**, 245–248 (2010).

4. M. Durot, P.-Y. Bourguignon, and V. Schachter, "Genome-scale models of bacterial metabolism: reconstruction and applications," FEMS microbiology reviews **33**, 164–190 (2009).

5. U. Consortium *et al.*, "Uniprot: the universal protein knowledgebase," Nucleic acids research **45**, D158–D169 (2017).

6. R. Overbeek, T. Begley, R. M. Butler, J. V. Choudhuri, H.-Y. Chuang, M. Cohoon, V. de Crécy-Lagard, N. Diaz, T. Disz, R. Edwards *et al.*, "The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes," Nucleic acids research **33**, 5691–5702 (2005).

7. D. Vallenet, E. Belda, A. Calteau, S. Cruveiller, S. Engelen, A. Lajus, F. Le Fèvre, C. Longin, D. Mornico, D. Roche *et al.*, "Microscopean integrated microbial resource for the curation and comparative analysis of genomic and metabolic data," Nucleic acids research **41**, D636–D647 (2013).

8. D. Vallenet, A. Calteau, S. Cruveiller, M. Gachet, A. Lajus, A. Josso, J. Mercier, A. Renaux, J. Rollin, Z. Rouy *et al.*, "Microscope in 2017: an expanding and evolving integrated resource for community expertise of microbial genomes," Nucleic acids research p. gkw1101 (2016).

9. I.-M. A. Chen, V. M. Markowitz, K. Palaniappan, E. Szeto, K. Chu, J. Huang, A. Ratner, M. Pillay, M. Hadjithomas, M. Huntemann *et al.*, "Supporting community annotation and user collaboration in the integrated microbial genomes (img) system," BMC genomics **17**, 307 (2016).

10. D. H. Haft, J. D. Selengut, L. M. Brinkac, N. Zafar, and O. White, "Genome properties: a system for the investigation of prokaryotic genetic content for microbiology, genome annotation and comparative genomics," Bioinformatics **21**, 293–306 (2005).

11. D. H. Haft, J. D. Selengut, R. A. Richter, D. Harkins, M. K. Basu, and E. Beck, "Tigrfams and genome properties in 2013," Nucleic acids research **41**, D387–D395 (2013).

12. I.-M. A. Chen, V. M. Markowitz, K. Chu, I. Anderson, K. Mavromatis, N. C. Kyrpides, and N. N. Ivanova, "Improving microbial genome annotations in an integrated database context," PLoS One **8**, e54859 (2013).

13. T. Lima, A. H. Auchincloss, E. Coudert, G. Keller, K. Michoud, C. Rivoire, V. Bulliard, E. De Castro, C. Lachaize, D. Baratin *et al.*, "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 (2009).

14. J. F. Sowa, *Knowledge representation: logical, philosophical, and computational foundations* (Course Technology, 1999).

15. Y. Shramko and H. Wansing, "Some useful 16-valued logics: How a computer network should think," Journal of Philosophical Logic **34**, 121–153 (2005).

16. Y. Shramko and H. Wansing, "Hyper-contradictions, generalized truth values and logics of truth and falsehood," Journal of Logic, Language and Information **15**, 403–424 (2006).

17. A. Morgat, E. Coissac, E. Coudert, K. B. Axelsen, G. Keller, A. Bairoch, A. Bridge, L. Bougueleret, I. Xenarios, and A. Viari, "Unipathway: a resource for the exploration and annotation of metabolic pathways," Nucleic acids research p. gkr1023 (2011).

18. R. D. Finn, P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas *et al.*, "The pfam protein families database: towards a more sustainable future," Nucleic acids research **44**, D279–D285 (2016).

19. R. Caspi, T. Altman, R. Billington, K. Dreher, H. Foerster, C. A. Fulcher, T. A. Holland, I. M. Keseler, A. Kothari, A. Kubo *et al.*, "The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases," Nucleic acids research **42**, D459–D471 (2014).

20. A. Morgat, T. Lombardot, K. B. Axelsen, L. Aimo, A. Niknejad, N. Hyka-Nouspikel, E. Coudert, M. Pozzato, M. Pagni, S. Moretti *et al.*, "Updates in rhea an expert curated resource of biochemical reactions," Nucleic acids research p. gkw990 (2016).

21. L. A. Vaas, J. Sikorski, B. Hofner, A. Fiebig, N. Buddruhs, H.-P. Klenk, and M. Göker, "opm: an r package for analysing omnilog® phenotype microarray data," Bioinformatics **29**, 1823–1824 (2013).

GROOLS: reactive graph reasoning for genome annotation through biological processes - Supplementary Data 1

Prediction accuracy for pathways (PWY) and functional units (FU) in 14 organisms

Organisms (UniProtKB proteome)		Genome Properties: UniProtKB annotations, falsehood mode					UniPathway: MicroScope annotaions, specific mode				
		TP	TN	FP	FN	ACC	TP	TN	FP	FN	ACC
Acinetobacter baylyi ADP1 (UP000000430)	PWY	13	8	4	2	77.78%	21	10	5	3	79.49%
	FU	75	104	14	3	91.33%	272	35	0	7	97.77%
Pseudomonas putida KT2440 (UP000000556)	PWY	17	6	4	2	79.31%	28	9	5	10	71.15%
	FU	96	55	6	3	94.38%	222	35	2	19	92.45%
Escherichia coli K-12 (UP000000625)	PWY	18	4	5	1	78.57%	27	6	2	9	75.00%
	FU	138	47	10	1	94.39%	111	17	5	15	86.49%
Delftia acidovorans DSM14801 (UP000000784)	PWY	16	6	3	2	81.48%	27	10	4	5	80.43%
	FU	105	35	3	7	93.33%	96	34	7	25	80.25%
Agrobacterium tumefaciens C58 (UP000000813)	PWY	17	3	3	6	68.97%	25	2	2	12	65.85%
	FU	100	25	6	56	66.84%	146	3	2	37	79.26%
Pedobacter heparinus DSM2366 (UP000000852)	PWY	8	7	1	9	60.00%	18	12	8	6	68.18%
	FU	62	51	2	103	51.83%	68	51	19	32	70.00%
Roseobacter litoralis DSM6996 (UP000001353)	PWY	11	3	3	4	66.67%	20	4	3	1	85.71%
	FU	69	29	3	10	88.29%	77	9	6	16	79.63%
Bacillus subtilis 168 (UP000001570)	PWY	18	4	3	1	84.62%	23	6	5	4	76.32%
	FU	130	33	10	1	93.68%	296	26	0	18	94.71%
Sphingomonas wittichii DSM6014 (UP000001989)	PWY	12	4	6	5	59.26%	22	11	9	5	70.21%
	FU	85	94	9	12	89.50%	87	59	14	12	84.88%
Spirosoma linguale DSM74 (UP000002028)	PWY	10	6	4	8	57.14%	18	11	5	4	76.32%
	FU	64	54	6	84	56.73%	66	36	9	12	82.93%
Chitinophaga pinensis DSM2588 (UP000002215)	PWY	9	4	4	8	52.00%	18	12	10	5	66.67%
	FU	62	40	15	49	61.45%	69	51	26	28	68.97%
Pseudomonas aeruginosa PAO1 (UP000002438)	PWY	18	3	3	4	75.00%	27	8	3	5	81.40%
	FU	130	29	3	5	95.21%	108	27	5	17	85.99%
Kytococcus sedentarius DSM20547 (UP000006666)	PWY	3	5	5	12	32.00%	15	16	7	9	65.96%
	FU	39	90	14	61	63.24%	47	69	19	44	64.80%
Salmonella typhimurium SL1344 (UP000008962)	PWY	14	3	5	1	73.91%	21	13	9	2	75.56%
	FU	81	24	16	1	86.07%	77	52	26	1	82.69%