Knowledge-based generalization of metabolic models

Anna Zhukova*
Inria Bordeaux Sud-Ouest/University of Bordeaux/CNRS UMR 5800
joint project-team Magnome
351, cours de la Libération F-33405 Talence cedex, France

David James Sherman†
Inria Bordeaux Sud-Ouest/University of Bordeaux/CNRS UMR 5800
joint project-team Magnome
351, cours de la Libération F-33405 Talence cedex, France

*corresponding author - anna.zhukova@inria.fr
†david.sherman@inria.fr

Abstract

Background: Genome-scale metabolic model reconstruction is a complicated process including (semi-)automatic inference of reactions participating in the organism's metabolism, followed by many iterations of network analysis and improvement. Better and better automatic model inference and analysis tools are being developed, but they may still miss some reactions or add erroneous ones. That is why the human expert's analysis of the model plays an important role at all the iterations of the reconstruction process. However, the size of the genome-scale models (i.e. thousands of reactions) makes it hard for a human to analyse them.

Results: To aid a human expert in metabolic model analysis we have developed a method for knowledge-based generalization that provides a higher-level view of a metabolic model, by masking inessential details while preserving its essential structure. The method groups biochemical species in the model into semantically equivalent classes and generalizes them into their common parent in the ChEBI ontology. The reactions between the same generalized species are factored together into generalized reactions.

Conclusions: We have applied our method to several metabolic models and shown that it improves understanding by helping to identify the peculiarities and potential errors, as well as facilitates model understanding and comparison.

Keywords: metabolic modelling; generalization; genome-scale reconstruction.

Introduction

Genome-scale metabolic models for new organisms include thousands of reactions. In most cases these reactions are automatically inferred by methods that combine databases of reactions and pathways with genomic information and existing models for similar organisms (Swainston et al., 2011). Genomic data for the new organism is compared to the data of the reference organism, to find genomic evidence such as the presence of catalysing enzymes for the reactions conserved in the new organism. Starting from the inference of a draft model, the model refinement process includes several iterations of model analysis, error detection, and improvement (Thiele and Palsson, 2010). The models produced at each iteration are intended for computer simulation, and so describe all the reactions thought to participate in the organism's metabolism. Although automatic model inference tools and genome comparison methods are becoming more and more advanced, they still may leave gaps in the model or add erroneous reactions. Thus, model evaluation by human experts remains important at all the iteration steps. However, because of their completeness, genome-scale models are too detailed and complicated to be easily understood by a human. The abundance of reactions in the model may hide errors.

For example, if in a genome-scale model of an yeast *Yarrowia lypolitica* (*MODEL1111190000* (Loira et al., 2012)) the enzyme *EC 2.3.1.16* were missing, a whole group of *Acyl-CoA:acetyl-CoA C-acyltransferase* reactions participating in the *Beta-oxidation of fatty acids* pathway (Metzler, 2001) would be eliminated: one for each of the six *3-oxoacyl-CoA* species presented in the model. However, the absence of these six reactions would be hidden by the other 59 reactions in the constitutive peroxisome of *Yarrowia lypolitica*, and a human expert may have difficulty noticing the error.

To aid human understanding of these complete models, a fair amount of work has been done on dividing them into reusable modules. Examples of such modules at the highest level is separating reactions in the model into compartment they are happening in.

Lower-level approaches can be divided into two groups: series ones and parallel. The series approaches operate with chains of reactions, and generalize them as a series, thus hiding the structure of the network. An example of a series approach is representing the network as a set of metabolic pathways (KEGG(Kanehisa et al., 2012), MetaCyC(Caspi et al., 2012)), that can be further divided, for example, into reaction modules (conserved sequences of reactions along the metabolic pathways)(Muto et al., 2013).

The other type of approaches operates with reactions that are "parallel", thus keeping the steps and preserving the general view of the network. An example of this approach is grouping reactions based on EC (Enzyme Commission) numbers(Tohsato et al., 2000). The drawback of this approach is that it is not applicable to networks with no EC number assigned or reactions with no catalysing enzymes identified. We developed another "parallel"-reaction method for knowledge-based generalization of metabolic models, that does not depend on enzyme information. It provides a higher-level view of a model while keeping its essential structure and omitting the details.

Definition 1 *The* model generalization *process groups chemical species present in the model into equivalence classes, and merges them into a generalized chemical species. Reactions that involve same generalized chemical species are then factored together into a generalized reaction.*

By applying the model generalization process, we can build a simplified model that focusses on the high level relationships. The simplified model can be further divided into pathways.

Mathematical basis

Basic definitions

We represent a *metabolic model* M as a pair of two sets: a set S of biochemical species, and a set R of reactions between them:

$$\begin{array}{rcll} M & = & \langle S, R\rangle & \text{- model,} \\ S & = & \{s_1, \ldots, s_n\} & \text{- species set,} \\ R & = & \{r_1, \ldots, r_m\} & \text{- reaction set.} \end{array}$$

We represent each *reaction* $r \in R$ as a pair of species sets: a set of its reactant species, and a set of its product species. A chemical reaction may be represented by a balanced chemical equation, showing the formulae of the reactants and products, and the changes that take place(Clugston and Flemming, 2000). This definition leads to the restriction (1) that all the species participating in the reaction should be different.

$$\begin{array}{rl} r = & \langle\{s_1^{(rs)}, \ldots, s_k^{(rs)}\}, \{s_1^{(ps)}, \ldots, s_l^{(ps)}\}\rangle \in R \subset \langle 2^S \times 2^S\rangle, \\ & \text{where } s_1^{(rs)} \neq \ldots \neq s_k^{(rs)} \neq s_1^{(ps)} \neq \ldots \neq s_l^{(ps)} \end{array} \qquad (1)$$

To perform the model generalization, we will define an *equivalence operation* $\sim$ on the species set, and group species into equivalence classes: $[s_i]^\sim = \{s_j \in S | s_j \sim s_i\}$.

Species equivalence imposes reaction equivalence: two reactions are equivalent if their corresponding reactant and product species are pairwise equivalent.

$$\begin{array}{rll} \forall r, \tilde{r} \in R & r = \langle\{s_1^{(rs)}, \ldots, s_k^{(rs)}\}, \{s_1^{(ps)}, \ldots, s_l^{(ps)}\}\rangle, & \\ & \tilde{r} = \langle\{\tilde{s}_1^{(rs)}, \ldots, \tilde{s}_{\tilde{k}}^{(rs)}\}, \{\tilde{s}_1^{(ps)}, \ldots, \tilde{s}_{\tilde{l}}^{(ps)}\}\rangle & \\ & k = \tilde{k}, l = \tilde{l}, & \\ r \sim \tilde{r} \iff & \wedge\ \forall i\ 0 \leq i \leq k\ \exists!\tilde{i}\ 0 \leq \tilde{i} \leq \tilde{k} :\ s_i^{(rs)} \sim \tilde{s}_{\tilde{i}}^{(rs)}, & \\ & \forall j\ 0 \leq j \leq l\ \exists!\tilde{j}\ 0 \leq \tilde{j} \leq \tilde{l} :\ s_j^{(ps)} \sim \tilde{s}_{\tilde{j}}^{(ps)}. & \end{array}$$

Equivalent reactions are factored together into a generalized reaction that operates with generalized species (i.e. species equivalence classes): $[r]^\sim = \langle\{[s_1^{(rs)}]^\sim, \ldots, [s_k^{(rs)}]^\sim\}, \{[s_1^{(ps)}]^\sim, \ldots, [s_l^{(ps)}]^\sim\}\rangle$.

In order to keep the number of distinct species participating in a reaction, the restriction (1'), analogous to the restriction (1), must be satisfied:

$$[s_1^{(rs)}]^\sim \neq \ldots \neq [s_k^{(rs)}]^\sim \neq [s_1^{(ps)}]^\sim \neq \ldots \neq [s_l^{(ps)}]^\sim \qquad (1')$$

In order to avoid creation of paths in the generalized model, that are not based on the evidence from the initial model, we introduce the restriction (2): species that do not participate in any pair of equivalent reactions and do not have any common equivalent species can not be grouped together (as there is no evidence of their equivalence in the initial model):

$$\begin{array}{rll} \forall s, \tilde{s} \in S\ s \sim \tilde{s}\ \Rightarrow\ \vee & \exists r \sim \tilde{r} \in R : s \in \mathit{reactants}(r) \wedge \tilde{s} \in \mathit{reactants}(\tilde{r}) & \\ & \exists r \sim \tilde{r} \in R : s \in \mathit{products}(r) \wedge \tilde{s} \in \mathit{products}(\tilde{r}) & (2) \\ & \exists \dot{s} \in S :\ s \tilde{\sim} \dot{s} \wedge \dot{s} \tilde{\sim} \tilde{s}. & \end{array}$$

Note, that restriction (2) can be reformulated as maximizing the number of species equivalence classes while keeping the reaction equivalence classes unchanged.

The *generalized model* $M/\sim$ is a pair of generalized species and reaction sets (quotient sets):

$$\begin{array}{rcll} M/\sim & = & \langle S/\sim, R/\sim\rangle & \text{- generalized model,} \\ S/\sim & = & \{[s_1]^\sim, \ldots, [s_{\tilde{n}}]^\sim\} & \text{- quotient species set,} \\ R/\sim & = & \{[r_1]^\sim, \ldots, [r_{\tilde{m}}]^\sim\} & \text{- quotient reaction set.} \end{array}$$

The generalized model is a "zoom out" of the initial model. It provides a higher-level view by including less species and reactions, but more generic ones. For example, *3-oxodecanoyl-CoA*, *3-oxolauroyl-CoA*, *3-oxohexanoyl-CoA*, and *3-oxooctanoyl-CoA* species of the initial model can be "zoomed out" into *oxo-fatty acyl-CoA* in the generalized model.

Every reaction of the generalized model corresponds to at least one reaction of the initial model having the same topology (number of distinct reactant and product species) and operating with species that can be "zoomed out" into those participating in the generalized reaction. The appropriate level of abstraction is defined by the initial model as the most general one that satisfies restrictions (1') and (2).

Specific and ubiquitous species

We say that a *ubiquitous species* is one that participates in many reactions (e.g. more than a threshold), such as *water*, *hydrogen*, *oxygen*, etc. Grouping species increases the number of reactions they participate in, while these are already shared by many reactions and common to most of the models. During visualisation these species are often even duplicated to improve readability (Rohn et al., 2012). That is why we do not generalize them. In the generalized model each of them forms a trivial equivalence class:

$$S^{(ub)} = \{s_1^{(ub)}, \dots, s_{\breve{n}}^{(ub)}\} \subset S : \forall i \, [s_i^{(ub)}]^{\sim} = \{s_i^{ub}\}$$

Specific species are all the others, they are divided into non-trivial equivalence classes and generalized accordingly.

Model generalization problem

Problem 1 *Given a metabolic model $M = \langle S, S^{(ub)} \subset S, R\rangle$ that describes n species (including $\breve{n} \leq n$ ubiquitous ones) and m reactions, find an equivalence operation $\sim$ that obeys the restrictions (1') and (2), and minimizes the number of reaction equivalence classes $\sharp R/\sim$.*

We will solve the model generalization problem 1 in three steps:

1. Define the most general equivalence operation $\mathring{\sim}$ (having minimal number of species equivalence classes $\sharp S/\mathring{\sim}$), that does not take into account the restrictions;
2. Modify the current equivalence operation, so that it satisfies the restriction (1');
3. Modify the current equivalence operation, so that it satisfies the restriction (2);

Step 1. Equivalence operation $\mathring{\sim}$.

Definition 2 *Given a model $M = \langle S, S^{(ub)} \subset S, R\rangle : \sharp S = n, \sharp S^{(ub)} = \breve{n} \leq n, \sharp R = m$, let us define an* equivalence operation $\mathring{\sim}$ *on the species set S as forming $\breve{n} + 1$ equivalence classes in the quotient set $S/\mathring{\sim}$: one for each of the ubiquitous species, and one for all the other species:*

$$\begin{array}{ll} \forall s^{(ub)} \in S^{(ub)} & [s^{(ub)}]^{\mathring{\sim}} = \{s^{(ub)}\}, \\ \forall s, \tilde{s} \in S \backslash S^{(ub)} & [s]^{\mathring{\sim}} = [\tilde{s}]^{\mathring{\sim}} = S \backslash S^{(ub)}. \end{array}$$

Lemma 1 *For any equivalence operation $\sim$ on the model $M = \langle S, S^{(ub)} \subset S, R\rangle$, the corresponding quotient species set $S/\sim$ and quotient reaction set $R/\sim$ are partitions of respectively the quotient species set $S/\mathring{\sim}$ and the quotient reaction set $R/\mathring{\sim}$ induced by $\mathring{\sim}$:*

$$\begin{array}{c} \forall \text{ equivalence operation } \sim \text{ defined on } \langle S, S^{(ub)}, R\rangle \\ \forall s \in S \quad [s]^{\sim} \subset [s]^{\mathring{\sim}} \\ \forall r \in R \quad [r]^{\sim} \subset [r]^{\mathring{\sim}} \end{array}$$

Algorithm 1 - Computation of $\mathring{\sim}$

Algorithm: Compute$\mathring{\sim}$
Data: $M = \langle S, S^{(ub)} \subset S, R\rangle : \sharp S = n, \sharp S^{(ub)} = \breve{n} \leq n, \sharp R = m$ - metabolic model describing n species, $\breve{n}$ among them being ubiquitous, and m reactions.
Result: $\mathring{\sim}$ - equivalence operation described in Lemma 1,
$M/\mathring{\sim} = \langle S/\mathring{\sim}, S^{(ub)}/\mathring{\sim} \subset S/\mathring{\sim}, R/\mathring{\sim}\rangle$ - corresponding generalized model.

$S/\mathring{\sim} \leftarrow \emptyset$; // resultant quotient species set $S/\mathring{\sim} \subset 2^S$
$S^{(ub)}/\mathring{\sim} \leftarrow \emptyset$; // resultant quotient ubiquitous species set $S^{(ub)}/\mathring{\sim} \subset 2^{S^{(ub)}}$
$R/\mathring{\sim} \leftarrow \emptyset$; // resultant quotient reaction set $R/\mathring{\sim} \subset 2^R$
$\mathring{\sim} \leftarrow \emptyset$; // resultant equivalence operation $\mathring{\sim} : S \cup R \rightarrow S/\mathring{\sim} \cup R/\mathring{\sim}$

/* Generalize ubiquitous species */
for $s^{(ub)} \in S^{(ub)}$ **do**
  $[s^{(ub)}]^{\mathring{\sim}} \leftarrow \{s^{(ub)}\}$; // map $s^{(ub)}$ to its equivalence class
end
$S^{(ub)}/\mathring{\sim} \leftarrow \{[s^{(ub)}]^{\mathring{\sim}} | s^{(ub)} \in S^{(ub)}\}$;

/* Generalize specific species */
for $s \in S \backslash S^{(ub)}$ **do**
  $[s]^{\mathring{\sim}} \leftarrow S \backslash S^{(ub)}$; // map s to its equivalence class
end
$S/\mathring{\sim} \leftarrow S^{(ub)}/\mathring{\sim} \cup \{S \backslash S^{(ub)}\}$;

/* Generalize reactions */
// map a reaction to its generalized version that operates with generalized species
$gen \leftarrow \lambda r.\langle \mathring{\sim}(reactants(r)), \mathring{\sim}(products(r))\rangle$;
for $r \in R$ **do**
  $[r]^{\mathring{\sim}} \leftarrow \{\tilde{r} \in R | gen(\tilde{r}) = gen(r)\}$;
end
$R/\mathring{\sim} \leftarrow \{[r]^{\mathring{\sim}} | r \in R\}$;

return $\mathring{\sim}, \langle S/\mathring{\sim}, S^{(ub)}/\mathring{\sim}, R/\mathring{\sim}\rangle$

The Compute$\mathring{\sim}$ algorithm forms the equivalence classes for ubiquitous and then specific species as in Definition 2 and then computes the generalized reactions.

Step 2. Stoichiometry preserving property obedience

Problem 2 *Given an equivalence operation $\sim$ defined on a metabolic model $M = \langle S, S^{(ub)} \subset S, R\rangle$ find an equivalence operation $\breve{\sim}$ that obeys property (1') and induces a quotient species set $S/\breve{\sim}$ of minimal size $\sharp S/\breve{\sim}$, such that $S/\breve{\sim}$ is a partition of the quotient species set $S/\sim$ induced by $\sim$, i.e., $\forall s \in S\ [s]^{\breve{\sim}} \subset [s]^{\sim}$.*

Algorithm

We will start with the given equivalence operation $\sim^0 = \sim$, and iteratively improve it, until the stoichiometry preserving property (1') is obeyed. We will denote the equivalence operation obtained at the i-th iteration step as $\sim^i$.

At each iteration, if there exists a species equivalence class that violates the stoichiometry preserving

property (1'), i.e.:

$$\exists s \neq \tilde{s} \in S, r \in R : s \in species(r) \wedge \tilde{s} \in species(r) \wedge [s]^{\sim^i} = [\tilde{s}]^{\sim^i},$$

we will partition this species equivalence class $[s]^{\sim^i} = [\tilde{s}]^{\sim^i}$ into two: $[s]^{\sim^{i+1}} \vee [\tilde{s}]^{\sim^{i+1}} = [s]^{\sim^i} = [\tilde{s}]^{\sim^i}$ to form a new approximation $\sim^{i+1}$ of the equivalence operation. When no species equivalence class violating the stoichiometry preserving property (1') can be found, the current equivalence operation is returned as result.

As at each iteration one equivalence species class is partitioned, resolving one of the conflicts. In the worst case, the equality operation = (each species is equivalent only to itself) will be achieved. As it obeys the stoichiometry preserving property (1'), the process will terminate.

Algorithm 2 - Stoichiometry Preserving Property Obedience

Algorithm: PreserveStoichiometry

Data: $\sim$ - equivalence operation defined on a metabolic model $M = \langle S, S^{(ub)} \subset S, R \rangle$,
$M/\sim = \langle S/\sim, S^{(ub)}/\sim \subset S/\sim, R/\sim \rangle$ - corresponding generalized model.

Result: $\breve{\sim}$ - equivalence operation described in Problem 3,
$M/\breve{\sim} = \langle S/\breve{\sim}, S^{(ub)}/\breve{\sim} \subset S/\breve{\sim}, R/\breve{\sim} \rangle$ - corresponding generalized model.

$S/\breve{\sim} \leftarrow S/\sim$; // resultant quotient species set $S/\breve{\sim} \subset 2^S$
$S^{(ub)}/\breve{\sim} \leftarrow S^{(ub)}/\sim$; // resultant quotient ubiquitous species set $S^{(ub)}/\breve{\sim} \subset 2^{S^{(ub)}}$
$R/\breve{\sim} \leftarrow \emptyset$; // resultant quotient reaction set $R/\breve{\sim} \subset 2^R$
$\breve{\sim} \leftarrow \sim$; // resultant equivalence operation $\breve{\sim} : S \cup R \rightarrow S/\breve{\sim} \cup R/\breve{\sim}$

/* Partition quotient species that do not obey the stoichiometry preserving property (1') */
for $S^{(gen)} \in \{\tilde{S}^{(gen)} \in S/\breve{\sim} | \exists s \neq \tilde{s} \in \tilde{S}^{(gen)}, r \in R : s \in species(r) \wedge \tilde{s} \in species(r)\}$ **do**
    $\Pi = Partition(S^{(gen)})$;

    // Update $S/\breve{\sim}$
    $S/\breve{\sim} \leftarrow S/\breve{\sim} \backslash \{S^{(gen)}\}$;
    $S/\breve{\sim} \leftarrow S/\breve{\sim} \cup \Pi$;

    // Update $\breve{\sim}$
    for $\tilde{S}^{(gen)} \in \Pi$ **do**
        for $s \in \tilde{S}^{(gen)}$ **do**
            $[s]^{\breve{\sim}} \leftarrow \tilde{S}^{(gen)}$;
        end
    end
end

/* Generalize reactions */
// map a reaction to its generalized version that operates with generalized species
$gen \leftarrow \lambda r.\langle \breve{\sim}(reactants(r)), \breve{\sim}(products(r)) \rangle$;
for $r \in R$ **do**
    $[r]^{\breve{\sim}} \leftarrow \{\tilde{r} \in R | gen(\tilde{r}) = gen(r)\}$;
end
$R/\breve{\sim} \leftarrow \{\breve{\sim}(r) | r \in R\}$;

return $\breve{\sim}, \langle S/\breve{\sim}, S^{(ub)}/\breve{\sim}, R/\breve{\sim} \rangle$

We will now describe the species equivalence class partition.

Clique partition

Definition 3 *For a given a set of species and a set of reactions between them, we define a* species compatibility graph *as a simple undirected graph with vertices representing the species, and edges linking those of the species that do not participate in the same reaction (i.e., putting them into the same equivalence class does not violate the stoichiometry preserving property (1')).*

Note, that any set of species that can be put into the same equivalence class without violating the stoichiometry preserving property (1'), forms a clique in the species compatibility graph, i.e. a complete subgraph: for every pair of its vertices there exists an edge linking them. Thus, the problem of partition the species equivalence class into minimum number of classes, such that all of them obey the stoichiometry preserving property (1') is a clique partition problem.

Problem 3 (Clique partition) *Find the smallest number of cliques in a graph such that every vertex in the graph is represented in exactly one clique.*

Remark 1 *Clique partition problem is known to be NP-complete (Bhasker and Samad, 1991).*

Species ontology

In a species compatibility graph, there are usually a few edges missing, and multiple solutions of the clique partition problem exist. In order to make the choice of the species equivalence classes biologically meaningful, we will use an ontology that describes hierarchical *is_a* relationships (i.e. more specific - more general) between biochemical species. This ontology can be viewed as a directed acyclic graph, with nodes representing terms describing species, and edges representing hierarchical relationships between them. A term T is an ancestor of a term t if and only if there exists a path from t to T.

Definition 4 *A term t is a* model term *if it corresponds to a specific species in the metabolic model.*

We will assume that no two model terms are connected by a descendant-ancestor relationship in the ontology (Otherwise, we will mark the ancestor term ubiquitous):
$\forall t, T \in terms\ (\exists\, species(t), species(T) \in S \wedge t \in descendants(T) \Rightarrow t = T)$.

We will iteratively remove all the leaf terms that are not model terms from the ontology, so that all the model terms become leaves, and all the leaves become model terms.

For each species equivalence class that needs to be partitioned, we will first find the least common ancestor T of the ontological terms corresponding to its species. If the ontology allows for multiple inheritance, and there are several such least common ancestors, we will pick the first one. Then we will look among the T-th descendant terms for those that are compatible (to avoid multiple inheritance).

Definition 5 *Terms $t_1, \ldots, t_k$ are* compatible *if and only if their descendant model terms do not intersect: $t_1, \ldots, t_k \in descendants(T)$ are compatible $\iff$ $\forall i \neq j \in \{1, \ldots, k\}$ $descendants(t_i) \cap descendants(t_k) \cap leaves(T) = \emptyset$.*

Problem 4 *Given a term T, find a compatible term set of minimal size that covers all the T-th descendant leaf terms and satisfies the stoichiometry preserving property (1'):*

$$
\begin{array}{rll}
 & k = k_{min}, & \\
 & t_1, \ldots, t_k \text{ are compatible}, & \\
? \, t_1, \ldots, t_k \in descendants(T) : \ \wedge & leaves(T) \subset descendants(t_1) \cup \ldots \cup descendants(t_k), & \\
 & \forall i \neq j \in \{1, \ldots, k\} \quad \forall s \in species(t_i), \tilde{s} \in species(t_j) & \\
 & \qquad \forall r \in R \ \{s, \tilde{s}\} \not\subset species(r). & (1'')
\end{array}
$$

To do so, we will first exclude all the terms that violate the stoichiometry preserving property (5). We thus obtain an exact set cover problem. We say that a subset S covers its own elements.

Problem 5 (Set cover) *Given a set X and a collection of its finite subsets Ψ, such that $\bigcup_{S\in\Psi} S = X$, find a minimum-size subset $\Pi \subset \Psi$ whose members cover all of X: $\bigcup_{S\in\Pi} S = \bigcup_{S\in\Psi} S = X$.*

Remark 2 *Set cover is NP-complete (Karp, 1972).*

Problem 6 (Exact set cover) *As in problem **??**, except that here the sets that are used in the cover are not allowed to intersect.*

Remark 3 *Exact cover is NP-complete (Goldreich, 2008).*

Exact set cover applied to ontological terms

Each ontological term t defines a set $S(t)$ of its descendant leaf terms (including t if it is a leaf). The instance consists of a set X of all leaf descendants of the least common ancestor T of the model terms of interest, and a collection Ψ of all sets defined by T-th descendant terms, and their relative complements with respect to X: $\forall S \in \Psi\ X\backslash S \in \Psi$, excluding all the sets that violate the stoichiometry preserving property (5). We look for a minimum-size exact cover of X.

Note, that in this case an exact cover always exists, e.g. the one formed by all the leaf terms.

Choice of the ontology

We will assume that any term that violates property (5) is removed from the ontology. Note, that the term T is also removed.

If the ontology has no multiple inheritance, i.e. $\forall S, \tilde{S} \in \Psi\ S \cap \tilde{S} \neq \emptyset \Rightarrow S \subseteq \tilde{S} \vee \tilde{S} \subseteq S$, the problem becomes trivial: The set of the root terms forms the solution. The size of the solution though depends on the characteristics of the ontology, e.g. for a completely flat ontology (i.e., a graph with no edges) the solution will consist of singleton equivalence classes.

If the multiple inheritance is allowed, any $\Psi \subseteq 2^X$ becomes possible, and the problem becomes *NP*-complete.

We will use the ChEBI ontology (de Matos et al., 2010) of chemical compounds, the *de facto* a standard for species annotation in metabolic models. ChEBI consists of three main branches: *chemical entity*, *role*, and *subatomic particle*. The *chemical entity* branch describes terms useful for annotation of biochemical species in a metabolic model. As of ChEBI version 101, this branch contains 37693 terms, among which 29888 are leaves. ChEBI has multiple inheritance with average number of parents 1.4 per term. Average number of siblings is also 1.4 per term. Maximal depth in the *chemical entity* branch is 28, while the average one is 11.

The level of details in the ChEBI hierarchy is not uniform: some sub-branches are more developed than others, which makes equally specific terms to be placed unequally deep in the hierarchical tree. For example, both *hydrogen peroxide (CHEBI:16240)* and *decanoyl-CoA (CHEBI:28493)* terms describe precise chemical molecules; but *hydrogen peroxide* is only 5 terms away from the *chemical entity* in the ChEBI hierarchy, while *decanoyl-CoA* is 11 terms away.

Besides that, different types of classification are combined together in the hierarchical tree, leading to multiple inheritance. For example, in the *fatty-acid (CHEBI:35366)* sub-branch, several types of the classification are present, including

- classification based on the length of the carbon chain:
  - *short-chain fatty acid (CHEBI:26666)*: 2-4 carbons;
  - *medium-chain fatty acid (CHEBI:59554)*: 6-12 carbons;
  - *long-chain fatty acid (CHEBI:15904)*: 14-22 carbons;

  - *very long-chain fatty acid (CHEBI:27283)*: 24 -26 carbons;

- classification based on the presence of double bonds in the carbon chain:
  - *saturated fatty acid (CHEBI:26607)*: no double bonds;
  - *unsaturated fatty acid (CHEBI:27208)*: one or more double bonds;

- classification based on substituent groups:
  - *hydroxy fatty acid (CHEBI:24654)*: one or more hydroxy substituents;
  - *oxo fatty acid (CHEBI:59644)*: at least one aldehydic or ketonic group;
  - *etc.*

Moreover, it turns out that using only hierarchical relationships in the ChEBI ontology is not always enough. Examples show, that similar reactions can happen to the acid and the base in a conjugate acid-base pair. A conjugate acid-base pair is two species, one an acid and one a base, that differ from each other through the loss or gain of a proton (Stoker, 2012). For instance, in the Rhea database of chemical reactions (Alcántara et al., 2012), the *acyl-CoA oxidase (RHEA:28354)* reaction: *decanoyl-CoA* + *FAD* + *H*+ → *trans-dec-2-enoyl-CoA*+$FADH_2$ is found for both *decanoyl-CoA (CHEBI:28493)* and its conjugate base *decanoyl-CoA(4-) (CHEBI:61430)*. But hierarchically, these species are very far from each other in the ChEBI ontology: The least common ancestor of *decanoyl-CoA* and *decanoyl-CoA(4-)* is *molecular entity (CHEBI:23367)*, a direct child of the root *chemical entity*. To establish a conjugate acid-base pair correspondence in the ChEBI ontology not the hierarchical (*is_a*) but special *is_conjugate_base_of* / *is_conjugate_acid_of* relationships are used. To maximize the chances of a conjugate acid-base pair being in the same quotient species set, we will generalize the hierarchical relationship:

Definition 6 *Term t is a* generalized direct descendant/ancestor *of a term T if and only if t or a conjugate base/acid of t is a direct descendant/ancestor of T or of a conjugate base/acid of T.*

Definition 7 *Term t is a* generalized descendant/ancestor *of a term T if and only if t is a generalized direct descendant/ancestor of T or of any generalized descendant/ancestor of T.*

We will extend Ψ so that it has closure under the operation of relative complement: $\forall S, \tilde{S} \in \Psi \; S \backslash \tilde{S} \in \Psi$. This will allow for solving the set cover problem instead of the exact cover one: As Ψ is closed under the operation of complement intersection, we can obtain an exact set cover $\tilde{C}$ from any set cover $C = \{S_1, S_2, \ldots, S_m\}$ by replacing its elements with their relative complements with the previous elements of C: $\tilde{C} = \{S_1, S_2 \backslash S_1, \ldots, S_m \backslash \bigcup_{i=1}^{m-1} S_i\}$.

To approximate the solution of the set cover problem, we will use a greedy algorithm.

Greedy Algorithm

Among the available subset candidates $S_i \in \Psi$ we will pick the one of the largest size and add it to the resulting set cover Π. We will repeat this operation until all elements of X are covered.

Algorithm 3 - Greedy Set Cover Obedience

Algorithm: GreedySetCover
Data: X - set of interest,
$\Psi \subseteq 2^X$ - set of subsets of X
Result: $\Pi \subseteq \Psi$ - set cover of X

$\Pi \leftarrow \emptyset$; // resultant cover

while $X \neq \emptyset$ **do**
  // select $S \in \Psi$ that covers maximum elements of X
  $S^{(max)} \leftarrow max(\Psi, criterion = \lambda S.\sharp(S \cap X))$;

  $\Psi \leftarrow \Psi \backslash \{S^{(max)}\}$;
  $X \leftarrow X \backslash S^{(max)}$;
  $\Pi \leftarrow \Pi \cup \{S^{(max)}\}$;
end

return Π

Greedy set cover is a polynomial time approximation algorithm that achieves an approximation ratio of $H(\sharp X)$, where $H(n)$ is the n-th harmonic number: $H(n) = \sum_{i=1}^{n} \frac{1}{i} \leq \ln n + 1$ (Chvatal, 1979). It is the best-possible polynomial time approximation algorithm for set cover, under plausible complexity assumptions (Feige, 1998).

Step 3. Species equivalence class number maximization

Problem 7 *Given an equivalence operation $\sim$ defined on a metabolic model $M = \langle S, S^{(ub)} \subset S, R \rangle$, such that $\sim$ obeys restriction (1'), find an equivalence operation $\tilde{\sim}$ that obeys restriction (2) and does not change the reaction equivalence classes: $R/\sim = R/\tilde{\sim}$.*

Algorithm

To satisfy the restriction (2) we will associate each species s in the initial model with a pair of reaction equivalence classes sets in the quotient reaction set $R/\sim$: those induced by reactions where it participates as a reactant or as a product:
$s \to \langle R_s^{(rs)} = \{[r_1^{(rs)}]^{\sim}, \ldots, [r_o^{(rs)}]^{\sim}\}, R_s^{(ps)} = \{[r_1^{(ps)}]^{\sim}, \ldots, [r_t^{(ps)}]^{\sim}\} \rangle$.

Definition 8 *Given an equivalence operation $\sim$ defined on a metabolic model $M = \langle S, S^{(ub)} \subset S, R \rangle$, such that $\sim$ obeys restriction (1'), let us define an* equivalence operation $\tilde{\sim}$ *as forming a separate species equivalence class for each of the ubiquitous species, and putting $\sim$-equivalent specific species that intersect in their product or reactant reaction classes in the same equivalence class:*

$$\begin{aligned} \forall s^{(ub)} \in S^{(ub)}, s \in S \;\; s^{(ub)} \tilde{\sim} s \quad &\iff \quad s^{(ub)} = s, \\ \forall s, \tilde{s} \in S \backslash S^{(ub)} \;\; s \tilde{\sim} \tilde{s} \quad &\iff \quad \wedge \quad \begin{array}{l} s \sim \tilde{s} \\ (R_s^{(rs)} \cap R_{\tilde{s}}^{(rs)} \neq \emptyset) \vee (R_s^{(ps)} \cap R_{\tilde{s}}^{(ps)} \neq \emptyset) \vee (\exists \dot{s} \in S : \; s \tilde{\sim} \dot{s} \wedge \dot{s} \tilde{\sim} \tilde{s}). \end{array} \end{aligned}$$

Any further partition of the quotient species set would imply the partition of the quotient reaction set. Hence the number of species equivalence classes is maximal for the current number of reaction equivalence classes, and the restriction (2) is satisfied.

Algorithm 4 - Maximization of the Number of Species Equivalence Classes

Algorithm: Maximize
Data: $\sim$ - equivalence operation defined on a metabolic model $M = \langle S, S^{(ub)} \subset S, R \rangle$, $M/\sim = \langle S/\sim, S^{(ub)}/\sim \subset S/\sim, R/\sim \rangle$ - corresponding generalized model.
Result: $\tilde{\sim}$ - equivalence operation described in Problem 2, $M/\tilde{\sim} = \langle S/\tilde{\sim}, S^{(ub)}/\tilde{\sim} \subset S/\tilde{\sim}, R/\tilde{\sim} \rangle$ - corresponding generalized model.

$S/\tilde{\sim} \leftarrow \emptyset$; // resultant quotient species set $S/\tilde{\sim} \subset 2^S$
$S^{(ub)}/\tilde{\sim} \leftarrow S^{(ub)}/\sim$; // resultant quotient ubiquitous species set $S^{(ub)}/\tilde{\sim} \subset 2^{S^{(ub)}}$
$R/\tilde{\sim} \leftarrow R/\sim$; // resultant quotient reaction set $R/\tilde{\sim} \subset 2^R$
$\tilde{\sim} \leftarrow \sim$; // resultant equivalence operation $\tilde{\sim} : S \cup R \rightarrow S/\tilde{\sim} \cup R/\tilde{\sim}$

/* Update specific species generalization */

// Map a species to a set of its $\sim$-equivalent species
// that participate in $\sim$-equivalent reactions
$r\_sim \leftarrow \lambda s.\{\tilde{s} \sim s | \exists r, \tilde{r} \in R : s \in reactants(r) \wedge \tilde{s} \in reactants(\tilde{r}) \wedge r \sim \tilde{r}\}$;
$p\_sim \leftarrow \lambda s.\{\tilde{s} \sim s | \exists r, \tilde{r} \in R : s \in products(r) \wedge \tilde{s} \in products(\tilde{r}) \wedge r \sim \tilde{r}\}$;
$sim \leftarrow \lambda s.r\_sim(s) \cup p\_sim(s)$;

$S/\tilde{\sim} \leftarrow S^{(ub)}/\tilde{\sim} \cup \{sim(s) | s \in S \backslash S^{(ub)}\}$;

// Merge all quotient species sets that intersect
while $\exists S^{(gen)} \neq \tilde{S}^{(gen)} \in S/\tilde{\sim} : S^{(gen)} \cap \tilde{S}^{(gen)} \neq \emptyset$ **do**
    $S/\tilde{\sim} \leftarrow (S/\tilde{\sim} \backslash \{S^{(gen)}, \tilde{S}^{(gen)}\}) \cup \{S^{(gen)} \cup \tilde{S}^{(gen)}\}$;
end

// Update $\tilde{\sim}$
for $S^{(gen)} \in S/\tilde{\sim}$ **do**
    for $s \in S^{(gen)}$ **do**
        $[s]^{\tilde{\sim}} \leftarrow S^{(gen)}$; // map s to its equivalence class
    end
end

return $\tilde{\sim}, \langle S/\tilde{\sim}, S^{(ub)}/\tilde{\sim}, R/\tilde{\sim} \rangle$

Complete Algorithm

As the most complex part of model generalization is the species partition, we will first do the other steps to minimize the size of each species quotient class to be partitioned. We will start with the equivalence operation $\mathring{\sim}$ described in Lemma 1, maximize the species equivalence class number for $\mathring{\sim}$, then obey the stoichiometry preserving property using the ChEBI ontology and greedy set cover algorithm, and finally maximize the species equivalence class number again.

Algorithm 5 - Computation of $\sim$

Algorithm: Compute$\sim$

Data: $M = \langle S, S^{(ub)} \subset S, R \rangle : \sharp S = n, \sharp S^{(ub)} = \breve{n} \leq n, \sharp R = m$ - metabolic model describing n species, $\breve{n}$ among them being ubiquitous, and m reactions.

Result: $\sim$ - approximation of the equivalence operation described in Problem 0, $M/\sim = \langle S/\sim, S^{(ub)}/\sim \subset S/\sim, R/\sim \rangle$ - corresponding generalized model.

$\mathring{\sim}, M/\mathring{\sim} \leftarrow Compute\mathring{\sim}(M)$;
$\tilde{\sim}, M/\tilde{\sim} \leftarrow Maximize(\mathring{\sim}, M/\mathring{\sim})$;
$\breve{\sim}, M/\breve{\sim} \leftarrow PreserveStoichiometry(\tilde{\sim}, M/\tilde{\sim})$;
$\sim, M/\sim \leftarrow Maximize(\breve{\sim}, M/\breve{\sim})$;

return $\sim, M/\sim = \langle S/\sim, S^{(ub)}/\sim, R/\sim \rangle$

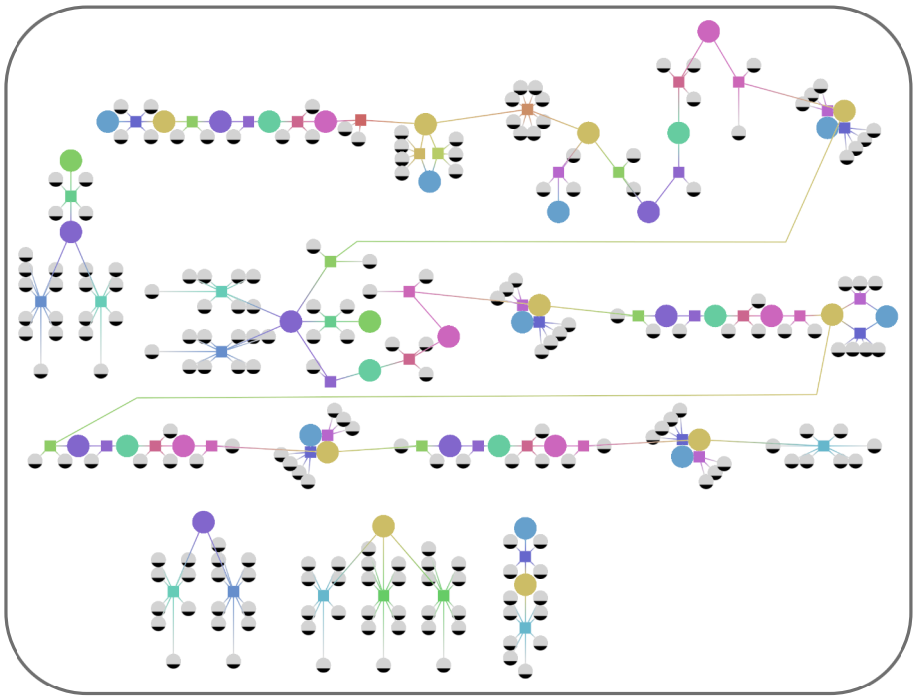
Applications

We have applied our method to three metabolic models that describe the *β-oxidation of fatty acids* pathway: a genome-scale metabolic model of the yeast *Yarrowia lipolytica* (*MODEL1111190000*), and two path2model (Li et al., 2010) pathways: fatty acid metabolism of the bacteria *Escherichia coli* (*BMID000000083160*) and of the yeast *Saccharomyces cerevisiae* (*BMID000000089673*). We have generalized these three models, and compared the results.

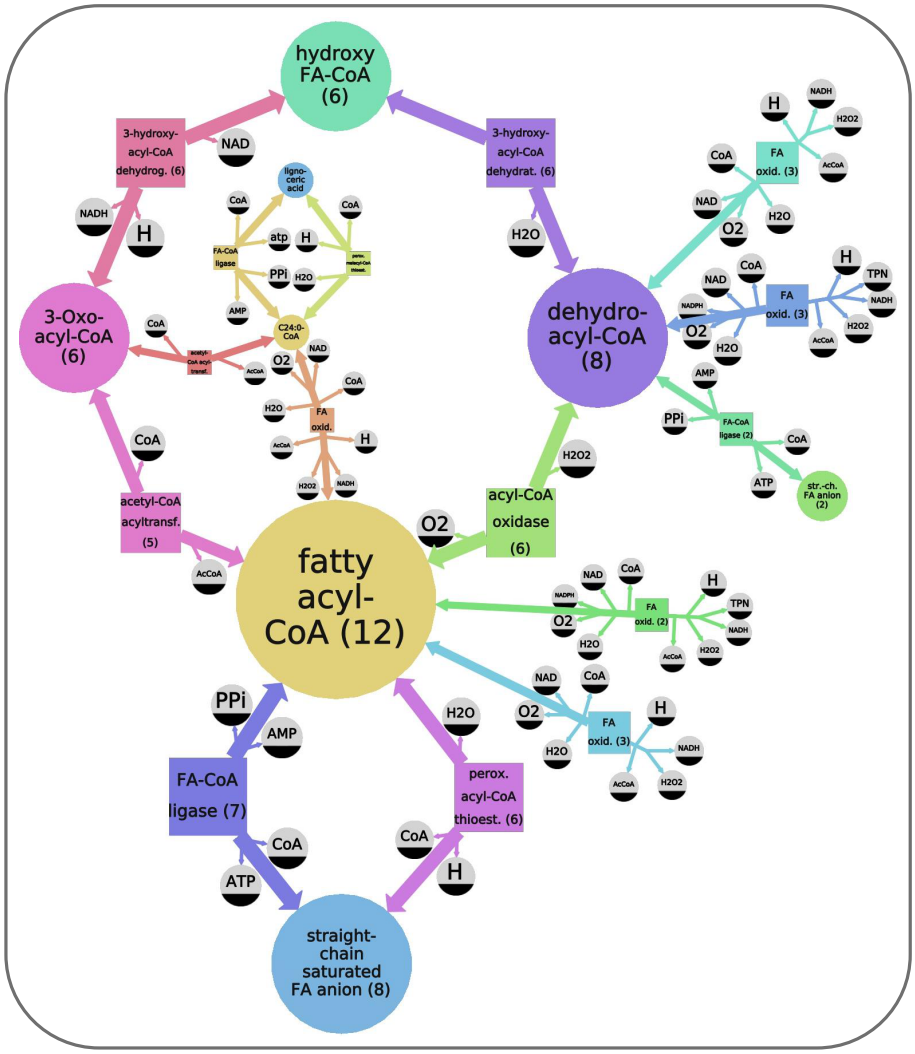
In *Yarrowia lypolitica* fatty acid oxidation happens in the peroxisome compartment, so we have extracted a sub-model that includes only those species and reactions that occur in the peroxisome (additional file 1).

The *Yarrowia lypolitica* model before and after generalization is represented in the figures 1 and 2.

Figure 1: *Yarrowia lypolitica* model



The figure shows *Yarrowia lypolitica* fatty acid oxidation model before generalization. Species are represented as circular nodes, and the reactions as squared ones, connected by edges to their reactants/products. Ubiquitous species are of smaller size and coloured gray. Specific species are divided into six equivalence classes, and coloured accordingly. Reactions are divided into fifteen equivalence classes, also represented by different colours.
The size of the model does not allow for readability of the species labels, thus we do not show them.
(The figure is produced using the Tulip (Auber, 2004) graph visualization tool.)



Generalization of the *Yarrowia lypolitica* model

Figure 2 *(previous page)*: The figure shows the generalization of the *Yarrowia lypolitica* fatty acid oxidation model (see figure 1). This model operates with quotient species and reactions. The number given in parenthesis and the size of each node indicates how many entities it generalizes. For example, the violet *dehydroacyl-CoA (8)* node is a quotient of 8 species: *hexadec-2-enoyl-CoA*, *oleoyl-CoA*, *tetradecenoyl-CoA*, *trans-dec-2-enoyl-CoA*, *trans-dodec-2-enoyl-CoA*, *trans-hexacos-2-enoyl-CoA*, *trans-octadec-2-enoyl-CoA*, and *trans-tetradec-2-enoyl-CoA* (coloured violet in figure 1). In a similar manner, the light-green *acyl-CoA oxidase (6)* quotient reaction, that converts *fatty acyl-CoA (12)* (yellow) into *dehydroacyl-CoA (8)* (violet), generalizes 6 corresponding light-green reactions of the initial model (figure 1).
The generalized model describes *β-oxidation* in a more generic way: as a transformation of *fatty acyl-CoA* (yellow) into *dehydroacyl-CoA* (violet), then into *hydroxyacy fatty acyl-CoA* (dark green), *3-ketoacyl-CoA* (magenta), and back to *fatty acyl-CoA* (with a shorter carbon chain); while the specific model describes the same process in more details, specifying those reactions for each of the *fatty acyl-CoA* species present in the organisms' cell (e.g. *decanoyl-CoA*, *dodecanoyl-CoA*, etc.). That is why the *beta-oxidation* chain of the reactions in the initial model, transforming step-by-step the fatty-acyl-CoA with the longest carbon chain into the one with the shortest chain, in the generalized model appears as a cycle (generalizing all the *fatty-acyls-CoA* into one species, regardless the chain-length).
The more precise model is needed for simulation, while the more general one is clearer to a human, and reveals the main properties of the model. For example, the generalized model highlights the fact that there is a particularity concerning *C24:0-CoA (tetracosanoyl-CoA)* (yellow): there exists a "short-cut" reaction (orange), producing it directly from another *fatty acyl-CoA* (yellow), avoiding the usual four-reaction beta-oxidation chain, used for other *fatty acyls-CoA*.
(The figure is produced using the Tulip graph visualization tool.)

The generalized *β-oxidation of fatty acids* models of *Escherichia coli* and *Saccharomyces cerevisiae* are shown in figure .

Discussions

We have developed a method that provides a "zoomed-out" view of a metabolic model, that keeps its essential structure but hides the details.

We have implemented our method as a Python program, that is available for download from `https://team.inria.fr/magn` It takes an SBML model as an input, annotates its species with ChEBI terms (if the annotations are not present in the model) and generalizes it. It produces a new SBML file, containing the generalized model, as an output.

We have applied our method to three metabolic models describing *beta-oxidation of fatty acids* and have shown that it helps finding gaps, and peculiarities in the models, as well as compresses the information stored in the model, which can be used for model visualisation and model comparison.

Acknowledgements

The authors would like to thank Dr. Nicolas Le Novère for discussions on metabolic modelling and SBML, and Dr. Romain Bourqui and Dr. Antoine Lambert of the LaBRI MABioVis team for advice on graph layout.

AZ was supported by a CORDI-S doctoral fellowship from Inria.

Author Disclosure Statement

No competing financial interests exist.

References

Alcántara, R., Axelsen, K.B., Morgat, A., et al. 2012. Rhea–a manually curated resource of biochemical reactions. *Nucleic Acids Research* 40(Database issue), D754–60.

Auber, D. 2004. Tulip A Huge Graph Visualization Framework. *In* M. Jünger, P. Mutzel, G. Farin, H.C. Hege, D. Hoffman, C.R. Johnson, K. Polthier, and M. Rumpf, eds., *Graph Drawing Software*, Mathematics and Visualization, 105–126. Springer Berlin Heidelberg.

Bhasker, J. and Samad, T. 1991. The clique-partitioning problem. *Computers & Mathematics with Applications* 22(6), 1–11.

Caspi, R., Altman, T., Dreher, K., et al. 2012. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic acids research* 40(Database issue), D742–53.

Chvatal, V. 1979. A Greedy Heuristic for the Set-Covering Problem. *Mathematics of Operations Research* 4(3), 233–235.

Clugston, M. and Flemming, R. 2000. *Advanced Chemistry (Advanced Science)*. OUP Oxford.

de Matos, P., Alcántara, R., Dekker, A., et al. 2010. Chemical Entities of Biological Interest: an update. *Nucleic Acids Research* 38(suppl 1), D249–D254.

Feige, U. 1998. A threshold of ln n for approximating set cover. *Journal of the ACM* 45(4), 634–652.

Goldreich, O. 2008. *Computational Complexity: A Conceptual Perspective*. Cambridge University Press, Cambridge.

Hucka, M., Hoops, S., Keating, S.M., et al. 2008. Systems Biology Markup Language (SBML) Level 2: Structures and Facilities for Model Definitions.

Kanehisa, M., Goto, S., Sato, Y., et al. 2012. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* 40(0305-1048 (Linking)), D109–14.

Karp, R.M. 1972. Reducibility Among Combinatorial Problems. *In* R.E. Miller and J.W. Thatcher, eds., *Complexity of Computer Computations*, 85–103. Plenum Press.

Li, C., Donizelli, M., Rodriguez, N., et al. 2010. BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Systems Biology* 4, 92.

Loira, N., Dulermo, T., Nicaud, J.M., et al. 2012. A genome-scale metabolic model of the lipid-accumulating yeast Yarrowia lipolytica. *BMC Systems Biology* 6(1), 35.

Metzler, D.E. 2001. *Biochemistry: The Chemical Reactions of Living Cells*. No. v. 1 in Biochemistry: The Chemical Reactions of Living Cells. Elsevier Science.

Muto, A., Kotera, M., Tokimatsu, T., et al. 2013. Modular Architecture of Metabolic Pathways Revealed by Conserved Sequences of Reactions. *Journal of chemical information and modeling* .

Rohn, H., Junker, A., Hartmann, A., et al. 2012. VANTED v2: a framework for systems biology applications. *BMC systems biology* 6(1), 139.

Stoker, H.S. 2012. *General, Organic, and Biological Chemistry*. Textbooks Available with Cengage YouBook Series. Brooks/Cole.

Swainston, N., Smallbone, K., Mendes, P., et al. 2011. The SuBliMinaL Toolbox: automating steps in the reconstruction of metabolic networks. *Journal of integrative bioinformatics* 8(2), 186.

Thiele, I. and Palsson, B.O. 2010. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols* 5(1), 93–121.

Tohsato, Y., Matsuda, H., and Hashimoto, A. 2000. A multiple alignment algorithm for metabolic pathway analysis using enzyme hierarchy. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology* 8, 376–83.

Figure 3 - Generalization of the *Escherichia coli* and *Saccharomyces cerevisiae* models

The figure shows the generalizations of the *Escherichia coli* (left) and *Saccharomyces cerevisiae* (right) fatty acid oxidation models. In the generalized model of *Saccharomyces cerevisiae* fatty oxidation is not a cycle, as it is in the generalized models of *Escherichia coli* and of *Yarrowia lypolitica* (figure 2). This is caused by the missing reactions operating with *hydroxy fatty acyls-CoA* (green), present in the *Escherichia coli* model.
(The figure is produced using the Tulip graph visualization tool.)

Additional Files

Additional file 1 — The peroxisome compartment of the *Yarrowia lipolytica* (*MODEL1111190000*) model

A sub-model of the *Yarrowia lipolytica* (*MODEL1111190000*) model that includes only those species and reactions that occur in the peroxisome compartment. In SBML level 2 version 4(Hucka et al., 2008) format with ChEBI annotations.