Retrieval, alignment, and clustering of computational models based on semantic annotations

Appendix

Abstract

In this supplementary appendix, we first describe the libSBAnnotation, a system for interconnecting and exploiting information from various biological databases. The information is stored in an ontology and can be used to compute semantic similarities between Biological Concepts, model elements, and entire models. We develop a number of similarity measures and discuss how the statistic significance of the model similarities can be assessed. The similarity measures are evaluated with benchmark models representing predefined categories, which serve as a gold standard. As an application, we align kinetic models to a large metabolic network.

Contents

1 libSBAnnotation - a query engine for biochemical annotations

When defining similarities between biochemical annotations, a basic requirement is to resolve and relate biological names and identifiers from public web resources. For this purpose, we developed the library libSBAnnotation[1]. Upon first invocation, entries describing biochemical compounds, reactions, genes, cellular compartments, and organisms are collected from various web resources (see Table 1) and equivalent resource entries are internally represented by "Biological Concepts" (BCs), which are linked to the original identifiers.

In contrast to the existing tools libAnnotationSBML [10], Saint [8], and BridgeDB [11], libSBAnnotation does not only provide cross-linking information or internal relations of the different web resources, but processes this information and resolves inconsistencies, e.g. cross-references between entities that are described to a different level of detail. The Biological Concepts are organised in several hierarchies via "is_a" relationships and interconnected by various other relationships extracted from the original resources. The resulting ontology allows users to compare entries from different web resources and to connect model elements by chains of biological relations even if this information could not be drawn from the individual web resources themselves. By enumerating the possible paths between ontology elements and evaluating the relations along these paths, similarities between elements can be defined (see section 6). The speed of libSBAnnotation queries depends on the maximum path length set for detecting indirect element relations. If many web resources are captured, it may be necessary to reduce the maximal path length in order to make queries reasonably fast.

When starting libSBAnnotation for the first time, the user can choose the web resources to be supported and the level of detail to which information is extracted. A reasonable selection covers more than 95% of the annotations in BioModels Database and uses 1.1 GB of RAM. Since the maximal amount of data available would consume more than 15 GB, an additional RESTful web service, to be accessed from different programming languages, is provided at `www.semanticsbml.org`. Apart from the gain of computational speed by having local copies of the web resources, a further difference between libSBAnnotation and libAnnotationSBML or Saint is its general applicability beyond the scope of SBML models. Finally, a fuzzy name search allows users to retrieve BCs with a defined literal error tolerance.

2 Similarity measures for Biological Concepts and MIRIAM annotations

In order to define similarities between MIRIAM-compliant element annotations, a first requirement is to score the similarities between the referenced Biological Concepts. To this aim, we consider a similarity measure for ontology elements that has previously been developed for semantic text analysis and modify it to make it compatible with our ontology. Given the similarities between Biological Concepts and numerical scores for the biological qualifiers, we can then define similarities between MIRIAM-compliant annotations. Further below, these similarity measures will be used to define similarities between models, either by the vector-based approach described in the main article or using structured similarity measures as described in sections 3 and 4.

[1] Open source python code is freely available at sourceforge `http://sourceforge.net/projects/semanticsbml/`

Web resource	**Annotation for**	**Relations extracted**
NCBI Taxonomy	organisms	is_a
Gene Ontology	compartments, processes	is_a, negatively_regulates, part_of, positively_regulates, regulates
ChEBI	species	has_functional_parent, has_parent_hydride, has_part, has_role, is_a, is_conjugate_acid_of, is_conjugate_base_of, is_enantiomer_of, is_substituent_group_from, is_tautomer_of
KEGG Compound	species	
KEGG Drug	species	
KEGG Enzyme	species	
KEGG Reaction	reactions	
KEGG Genome	genes	
Reactome	species	
EntrezGene	genes	encodes, hasFunction, inOrganism, inProcess, isLocated, isPartOf
UniProt	species	encodes, hasProcess, inOrganism
Interpro	species	parent, member, example, found_in
Saccharomyces Genome Database	species	

Table 1: List of web resources supported by the libSBAnnotation. Upon installation, the web resources are screened for element names, identifiers, relations, and cross-linking information. Corresponding relations from different web resources (e.g., "is_a") are combined in a single ontology.

2.1 Mathematical notation

First of all, let us introduce some formal notation for models, model elements, their annotations, the referenced resource entries or Biological Concepts, and the relations between them (compare Figure 1).

Models and annotations. Formally, we identify a model M (from a model list $\mathcal{M}$) with a set of model elements $m \in M$, while each element is identified with a set of MIRIAM-compliant annotations $\mu^{\mathrm{A}} \in m$. Each annotation relates the model element to an identifier (ID) $\mu^{\mathtt{I}}$ from a web resource $\mu^{\mathtt{R}}$, while the qualifier $\mu^{\mathtt{Q}}$ specifies the relation between the element and the corresponding resource entry. Thus an annotation is formally a triple $\mu^{\mathrm{A}} = (\mu^{\mathtt{R}}, \mu^{\mathtt{I}}, \mu^{\mathtt{Q}})$. In addition, the libSBAnnotation links each known web resource entry $(\mu^{\mathtt{R}}, \mu^{\mathtt{I}})$ to a Biological Concept μ, a basic element of the ontology. While a BC can be associated with entries from several web resources, the mapping of an entry to a BC $\mu = \mathtt{BC}(\mu^{\mathtt{R}}, \mu^{\mathtt{I}})$ is unique. Therefore, whenever an annotation points to a resource element listed in the libSBAnnotation ontology, it can be represented by the pair $(\mu, \mu^{\mathtt{Q}})$.

Relations and properties of Biological Concepts In our ontology, BCs are connected by directed relation edges, described as triples $r = r(\mu, \nu, relation\_type)$ of the two related BCs and the relation type (for a list of relation types, see Table 2). Each relation implicitly defines an inverse relation $r(\nu, \mu, inverse\_relation\_type)$, where the inverse type is e.g. "part_of" for "has_part". The set of outgoing relation edges at BC μ is called $R(\mu)$ and the set of outgoing relation edges of a certain type is called $R(\mu, relation\_type)$. Each BC μ can be characterised by a number of properties. Its depth $d(\mu)$ is defined as the

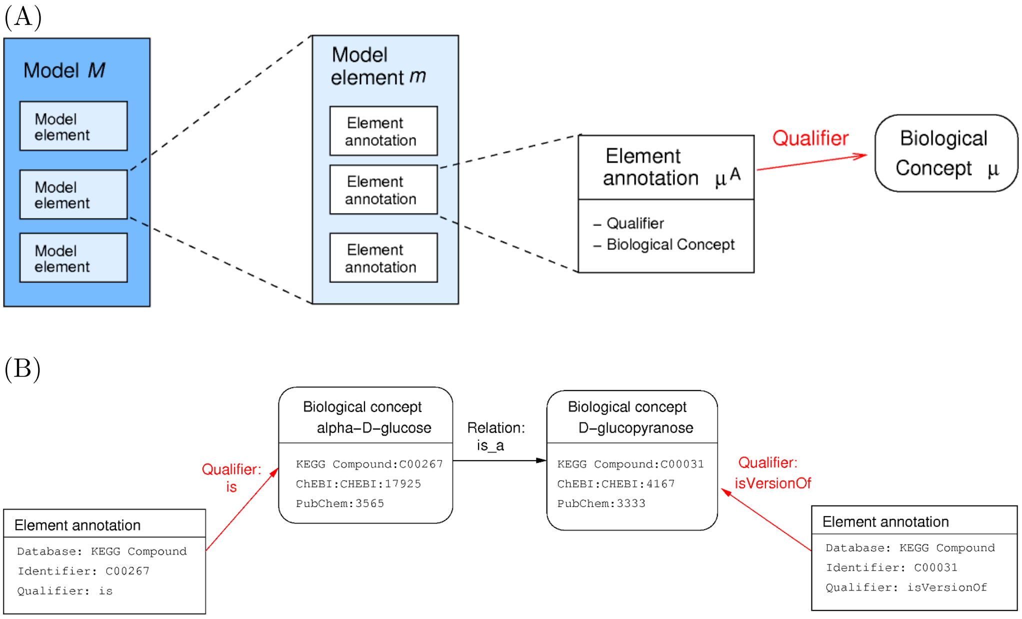


Figure 1: Models and element annotations. (A) Structure of an annotated SBML model. A model M (left) contains a number of elements. Each model element m can have multiple annotations and each annotation μ^{A} refers to a Biological Concept (BC) via a biological qualifier. (B) Semantic relation between element annotations. Two element annotations (left and right boxes) refer, via their biological qualifiers (red arrows), to Biological Concepts (α-D-glucose and D-glucopyranose). The two BCs are linked by biological relations represented in the libSBAnnotation (black arrows), which can either be direct (as shown in here) or consist of a chain of relations with other BCs in between.

length of the “is_a” relations path between the ontology element and its root. Similarly, the height $h(\mu)$ of a BC is given by the maximal length of an “is_a” relation path to any of the leaves below it. Finally, we define the frequency c_μ of a BC in BioModels Database, the cumulative frequency $c^c_\mu = c_\mu + \sum_{\xi: r(\mu,\xi,t) \in R(\mu, is\_a)} c^c_\xi$ including all its children in the “is_a” hierarchy, and the total number of annotation appearances including pseudocounts $c_\Omega = 1 + \sum_{\forall \xi: c_\xi > 0} (c_\xi + 1)$. To include MIRIAM annotations that do not match any known BC in our ontology, we create for each of them a new BC without any relations to other concepts.

2.2 Similarity measure $\sigma^{\mathrm{Li}}_{\mathrm{BC}}$ adopted from semantic text analysis

We now introduce quantitative similarities between Biological Concepts, based on their frequencies and the known relations between them. First, we adopt a similarity measure for terms in natural language, introduced by Li *et al.* [7], which is based on a large text corpus and on a semantic knowledge base with a simple hierarchical “is_a” relation. The similarity of two terms reflects three different factors: (i) their distance in the relation tree, (ii) the depth of their lowest common ancestor in the relation tree, and (iii) their count numbers in the text corpus. In the formula for similarities, these aspects are described by three multiplicative factors f_1, f_2, and f_3. If we simply replaced natural language terms by

relation type	f_{rts}
is_a	.5
part_of	.1
has_part	.1
regulates	.01
positively_regulates	.01
negatively_regulates	.01
is_tautomer_of	.9
is_enantiomer_of	.01
is_conjugate_acid_of	.9
is_conjugate_base_of	.9
has_role	.75
has_functional_parent	0.
is_substituent_group_from	.01
has_parent_hydride	.9
encodes	.1
hasFunction	.75
hasProcess	.25
inOrganism	.1
inProcess	.25
isLocated	.1
isPartOf	.25

Table 2: Quantitative factors f_{rts} assigned to the relations in the libSBAnnotation ontology. The numerical values were chosen ad-hoc after a series of tests and systematic evaluations.

Biological Concepts, we would obtain the similarity measure

$$\sigma_{\mathrm{BC}}^{\mathrm{Li}}(\mu,\nu) \quad = \quad f_1(\mu,\nu)\cdot f_2^{\mathrm{Li}}(\mu,\nu)\cdot f_3^{\mathrm{Li}}(\mu,\nu) \qquad (1)$$

with the formulae for f_1, f_2^{Li}, and f_3^{Li} proposed in [7]. In contrast to Li's original application, we deal with an ontology containing several relation types, so we had to modify these formulae for our purposes. First, the different types of relations have to be incorporated into the distance term. Second, because of the various additional relation types, the relationship graph may not be a tree, but may contain cycles and consist of disconnected subgraphs. We therefore developed new formulae for f_1, f_2, and f_3, but attempted to keep them similar to the ones suggested by Li *et al.*

Ontology distance factor f_1 The ontology distance factor f_1 accounts for the paths between two Biological Concepts in the ontology and for the types of relations along these paths. We score each relation arrow $r(\mu,\nu,t)$ of type t by a number $f_{\mathrm{rts}}(t)$ between 0 and 1 (see Table 2), multiply these scores along each possible path, and choose the path with the maximal resulting value. The resulting value f_1 can be recursively defined as

$$\begin{aligned} f_1(\mu,\mu) &= 1 \\ f_1(\mu,\nu) &= \max_{t,\xi:\, r(\mu,\xi,t)\in R(\mu)} \left(f_{\mathrm{rts}}(t)\cdot f_1(\xi,\nu)\right). \end{aligned} \qquad (2)$$

We further set $f_1(\mu,\nu)=0$ if there is no relation chain between μ and ν or if this chain is too long to be calculated in reasonable time. In practise, we only evaluate paths up to a certain maximal length.

Ontology depth factor f_2 Model annotations can describe the intended BCs to different levels of detail. For instance, a model element could either be annotated with the ChEBI entry CHEBI:18133 (for hexose) or with CHEBI:17925 (for α-D-glucose). Since "hexose" is more general than "α-D-glucose", two elements annotated as α-D-glucose will potentially be more similar than two elements annotated as hexoses. To capture this in the formula for BC similarities, a second factor f_2 is included into the similarity measure. It increases with the degree of detail of an annotation, which is measured by its depth in the ontology tree. In Li's original formula, the level of detail of the two BCs is determined by the depth of their lowest common ancestor. In our case, two compared BCs, e.g. a protein and the gene encoding it, are not necessarily part of the same tree of "is_a" relations and do not have a common ancestor. Therefore, we consider the average relative depth of both BCs and use the formula

$$f_2^{\mathrm{Li}}(\mu, \nu) \quad = \quad \tanh\left(\frac{3}{2}\left(\frac{d(\mu)+1}{d(\mu)+h(\mu)+1} + \frac{d(\nu)+1}{d(\nu)+h(\nu)+1}\right)\right). \tag{3}$$

The prefactor 3/2 was chosen ad-hoc to use the nonlinear range of the hyperbolic tangent function.

Local semantic density factor f_3 Some annotations, for instance the GeneOntology term for "cell" (GO:0005623), are very frequent in BioModels Database (see BioModelsStats website at `www.semanticsbml.org`) and provide little information to distinguish between models. Just like the unspecific annotations discussed before, such frequent annotations could be down-weighted in the overall model similarity. Since there is no "text corpus" for biochemical annotations, we use the collected annotations from the BioModels Database instead. We screen all models for references to each of the BCs and compute the local semantic density term

$$f_3^{\mathrm{Li}}(\mu, \nu) \quad = \quad \tanh\left(-\log\left(\frac{\min(c_\mu^c, c_\nu^c)}{c_\Omega}\right)\right). \tag{4}$$

from the cumulative frequencies c_μ^c (see definition above).

2.3 Similarity measure $\sigma_{\mathrm{BC}}^{\mathrm{DD}}$ accounting for distance/depth dependence

Li's formula Eq. (1) treats the distance and depths of different ontology terms as independent factors. In reality, if two BCs have a small depth, their maximum distance will be limited by this very fact. To avoid the independence assumption, we defined a new similarity measure for Biological Concepts in which distance and depth are combined and the local semantic density is calculated individually for each BC:

$$\sigma_{\mathrm{BC}}^{\mathrm{DD}}(\mu, \nu) = f_1(\mu, \nu)^{f_2(\mu,\nu)} \cdot f_3(\mu) \cdot f_3(\nu) \tag{5}$$

where

$$f_2(\mu, \nu) \quad = \quad \frac{2}{d(\mu)+d(\nu)+2} \tag{6}$$

$$f_3(\mu) \quad = \quad 1 - \frac{\log(c_\mu^c + 1)}{\log c_\Omega}. \tag{7}$$

2.4 Similarity measure σ_{An} for MIRIAM annotations

As shown in Figure 1 (B), two model elements are related by the Biological Concepts referenced in their annotations. Model elements and BCs are connected via qualifiers, while BCs may be interconnected by relation chains in the ontology. We quantify these relations by a similarity measure between element annotations μ^{A} and ν^{A}. The similarity should increase with the similarity between the two BCs and with direct qualifiers like "is", in contrast to vague qualifiers like "isVersionOf". Since these two factors are logically independent, we combine them by the multiplicative formula

$$\sigma_{\mathrm{An}}(\mu^{\mathrm{A}}, \nu^{\mathrm{A}}) = f_{\mathrm{qsm}}(\mu^{\mathtt{Q}}, \nu^{\mathtt{Q}}) \cdot \sigma_{\mathrm{BC}}(\mu, \nu). \tag{8}$$

The term f_{qsm} scores each possible pair of qualifiers by a value between 0 and 1 (for numerical values, see Table 3). The term $\sigma_{\mathrm{BC}} \in [0, 1]$ describes the similarity between BCs as determined from the ontology, i.e., either $\sigma_{\mathrm{BC}}^{\mathrm{Li}}$ from Eq. (1) or $\sigma_{\mathrm{BC}}^{\mathrm{DD}}$ from Eq. (5).

f_{qsm}	is	isDescribedBy	isVersionOf	hasVersion	isHomologTo	isPartOf	hasPart	isEncodedBy	encodes
is	1.	0.	.5	.5	.8	.2	.2	.2	.2
isDescribedBy	0.	1.	0.	0.	0.	0.	0.	0.	0.
isVersionOf	.5	0.	.3	.25	.4	.1	.1	.1	.1
hasVersion	.5	0.	.25	.3	.4	.1	.1	.1	.1
isHomologTo	.8	0.	.4	.4	.7	.64	.64	.64	.64
isPartOf	.2	0.	.1	.1	.64	.05	.04	.04	.04
hasPart	.2	0.	.1	.1	.64	.04	.05	.04	.04
isEncodedBy	.2	0.	.1	.1	.64	.04	.04	.5	.04
encodes	.2	0.	.1	.1	.64	.04	.04	.04	.5

Table 3: Contribution $f_{\mathrm{qsm}}(\mu^{\mathtt{Q}}, \nu^{\mathtt{Q}})$ of the biological qualifiers to the annotation similarity Eq. (8). Each possible pair of biological qualifiers $\mu^{\mathtt{Q}}$ and $\nu^{\mathtt{Q}}$ is scored by a value between 0 and 1. The numerical values were chosen ad-hoc after a series of tests.

3 Vector-based similarity measure $\sigma_{\mathrm{Mo}}^{\mathrm{TVSM}}$ for models

Given the pairwise similarities between element annotations, we can now define similarities between entire models. The first type of similarity measures, which is used in our online model search and discussed in the main article, is related to the Topic-based Vector Space Model (TVSM, [1]) used in information retrieval. The basic idea is to compare models by their feature vectors, i.e., the columns of the annotation matrix. In the feature vector v_{M} of model M, the i$^{\mathrm{th}}$ BC μ is represented by a component v_{iM} with a value of 1 if the model points to this BC by one of its annotations and $v_{iM} = 0$ otherwise. Instead of the value 1, one may also choose different values $v_{iM} = \sqrt{f_{\mathrm{qsm}}(\mu^{\mathtt{Q}}, \mu^{\mathtt{Q}})}$ depending on the qualifier appearing in the model. If a BC is referenced several times in a model, one could use either the sum or the maximum of the different values.

Simple similarities between two models M and N can be defined based on the scalar product $v_M \cdot v_N$ between feature vectors, which basically counts how many of the BCs are shared by both models. Using different normalisations, we can obtain our cosine coefficient $v_M \cdot v_N/\sqrt{||v_M||\,||v_N||}$, but also other association measures like Dice's coefficient $2\,v_M \cdot v_N/(|v_M|_1 + |v_N|_1)$ or the overlap coefficient $v_M \cdot v_N/\min(|v_M|_1, |v_N|_1)$, where we assumed binary feature vectors [12]. If two models contain BCs that are similar, but not identical, the cosine coefficient will not take into account their similarity. In the Topic-based Vector Space Model (TVSM) approach, we replace the scalar product by the quadratic form

$$\sigma_{\mathrm{Mo}}^{\mathrm{TVSM}}(M,N) = \frac{v_{\mathrm{M}}{}^{\mathrm{T}}\,S\,v_{\mathrm{N}}}{\sqrt{v_{\mathrm{M}}{}^{\mathrm{T}} S v_{\mathrm{M}}}\sqrt{v_{\mathrm{N}}{}^{\mathrm{T}} S v_{\mathrm{N}}}}, \tag{9}$$

where S is the precalculated similarity matrix between BCs. For instance, we could use the similarity measure $\sigma_{\mathrm{BC}}^{\mathrm{DD}}$ Eq.(5) and set $S_{il} = f_1(\mu,\nu)^{f_2(\mu,\nu)} \cdot f_3(\mu) \cdot f_3(\nu)$ where μ is the i^{th} and ν is the l^{th} BC. Since the feature vectors v_{M} and all individual individual components of S are non-negative, the resulting model similarities will be non-negative even if S is not a positive-definite matrix.

For positive-definite matrices S, this formula can be can be justified by the use of linearly transformed feature vectors $v'_{\mathrm{M}} = Av_{\mathrm{M}}$. The transformation is chosen such that $A^{\mathrm{T}}A = S$. Effectively, this reduces the angle between basis vectors representing similar BCs. The resulting similarity measure is defined as the cosine of the angle between the transformed vectors v'_{M} and v'_{N}.

4 Structured similarity measures for model elements and models

The vector-based similarity measures detect which BCs are referenced by a model, but not how they are linked to individual model elements. In SBML models, each element (describing a compound, reaction, etc.) can have multiple annotations with different qualifiers and referring to different BCs. To account for the specific arrangement of annotations, we developed a second class of model similarity measures, which are described in this section.

In these *structure-based* similarity measures, the similarity between two models is computed from the pairwise similarities between model elements, which in turn is based on the pairwise similarities between their individual annotations. We consider two different approaches: First, in the *preference-based* approach, we iterate over all model elements, determine for each of them the maximal similarity to elements from the other model, and average over the resulting values. Using the same approach, the similarities between two model elements can be computed from the pairwise similarities of the annotations. Second, we discuss an approach in which annotations are regarded as uncertain pieces of information and are combined using formulae derived from probabilistic reasoning.

4.1 Preference-based similarity $\sigma_{\mathrm{El}}^{\mathrm{Pref}}$ for model elements

The preference-based similarity between two model elements m and n can be computed from the pairwise similarities $\sigma_{\mathrm{An}}(\mu^{\mathrm{A}},\nu^{\mathrm{A}})$ of all their annotations μ^{A} and ν^{A} by the formula

$$\sigma_{\mathrm{El}}^{\mathrm{Pref}}(m,n) = \frac{\sum\limits_{\mu^{\mathrm{A}}\in m} \max\limits_{\nu^{\mathrm{A}}\in n} \sigma_{\mathrm{An}}(\mu^{\mathrm{A}},\nu^{\mathrm{A}}) + \sum\limits_{\nu^{\mathrm{A}}\in n} \max\limits_{\mu^{\mathrm{A}}\in m} \sigma_{\mathrm{An}}(\mu^{\mathrm{A}},\nu^{\mathrm{A}})}{|m| + |n|}, \tag{10}$$

where $|m|$ is the number of annotations assigned to model element m. For every annotation, the most similar annotation from the other model is selected; an example is shown in Figure

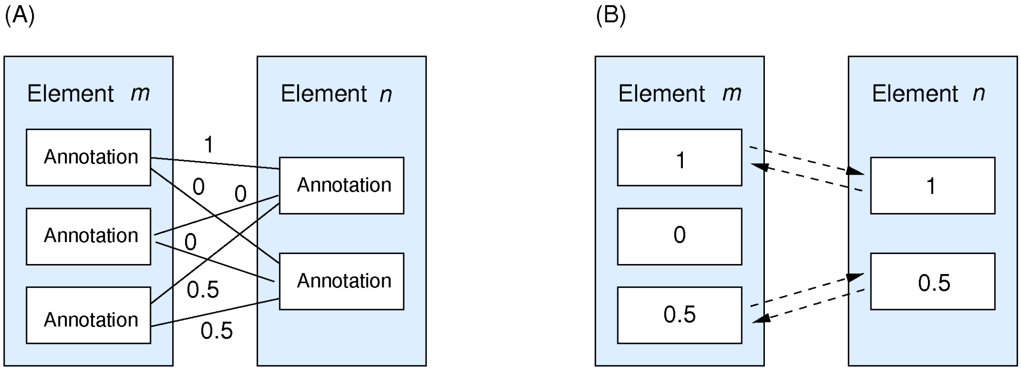


Figure 2: Preference-based similarity measure $\sigma_{\mathrm{El}}^{\mathrm{Pref}}$ for model elements. (A) The similarity between two model elements m and n (blue boxes) is computed from the pairwise similarities σ_{An} (shown as numbers) between their annotations (small white boxes). (B) For each annotation, the maximal similarity to an annotation from the other model element is determined (dotted arrows). To compute the overall element similarity $\sigma_{\mathrm{El}}^{\mathrm{Pref}}(\mu^{\mathrm{A}}, \nu^{\mathrm{A}})$, these numbers are averaged over all annotations, $\sigma_{\mathrm{El}}^{\mathrm{Pref}}(\mu^{\mathrm{A}}, \nu^{\mathrm{A}}) = (1 + 0 + 0.5 + 1 + 0.5)/5 = 0.6$.

2. If one of the elements has no annotations, the similarity $\sigma_{\mathrm{El}}^{\mathrm{Pref}}(m, n)$ is set to a small value $\varepsilon_{\mathrm{El}} \geq 0$, representing the similarity between randomly picked elements (see Table 4). Equation (10) yields similarity values between 0 and 1. A similarity of 1 shows that the annotations of both elements contain `is` qualifiers and are basically identical. A similarity of $0 < \varepsilon_{\mathrm{El}}$, on the contrary, shows that both elements are annotated, but their annotations are completely unrelated.

Parameter	**Value**
basal element similarity $\varepsilon_{\mathrm{El}}$	0
basal model similarity $\varepsilon_{\mathrm{Mo}}$	0

Table 4: Numerical parameters used on calculations for the preference-based model similarity measure.

4.2 Preference-based similarity $\sigma_{\mathrm{Mo}}^{\mathrm{Pref}}$ for models

Similarity between models Similarities between models could be computed from the pairwise similarities of their elements in a variety of ways. In analogy to formula (10), we propose the preference-based formula

$$\sigma_{\mathrm{Mo}}^{\mathrm{Pref}}(M, N) = \frac{\sum_{m \in M} \max_{n \in N} \sigma_{\mathrm{El}}(m, n) + \sum_{n \in N} \max_{m \in M} \sigma_{\mathrm{El}}(m, n)}{|M| + |N|}. \tag{11}$$

A model similarity of 1 means that each element from one model has a similarity of 1 with at least one element from the other model, while a model similarity of 0 means that any pair of elements has similarity 0. Again, if one of the models has no elements at all, the similarity $\sigma_{\mathrm{Mo}}^{\mathrm{Pref}}(M, N)$ is set to a value $\varepsilon_{\mathrm{Mo}} \geq 0$, representing a basal similarity between randomly picked models (see Table 4).

Normalised model similarities Our preference-based similarity measure can lead to counter-intuitive results if models are incompletely annotated or contain several elements with the same annotations. For instance, a model with missing annotations will have a self-similarity $\sigma_{\mathrm{Mo}}^{\mathrm{Pref}}(M, M)$ smaller than 1 and may have higher similarities to other, different models. To compensate for such artifacts caused by incomplete annotation, the similarity of two models can be normalised by the maximum of their self-similarities, yielding the normalised similarity

$$\hat{\sigma}_{\mathrm{Mo}}^{\mathrm{Pref}}(M, N) = \frac{\sigma_{\mathrm{Mo}}^{\mathrm{Pref}}(M, N)}{\max(\sigma_{\mathrm{Mo}}^{\mathrm{Pref}}(M, M), \sigma_{\mathrm{Mo}}^{\mathrm{Pref}}(N, N))}. \tag{12}$$

4.3 Similarities $\sigma_{\mathrm{El}}^{\mathrm{Prob}}$ derived from probabilistic reasoning

If a human expert had to compare model elements manually, she would go through the annotation lists, logically combine all pieces of information, and try to find pairs of annotations that either support or preclude similarity. Elements without annotations would be regarded as dissimilar since their odds to describe the same concepts are rather low. Finally, where annotations do not seem to fit at all, an expert might conclude that some of the annotations are wrong. The formula (10) for the preference-based element similarity $\sigma_{\mathrm{El}}^{\mathrm{Pref}}$ reflects this human reasoning only very roughly. To mimic a little better how human experts would compare model elements, we developed another similarity measure based on ideas from probabilistic reasoning. Although this measure performed worse than heuristic ones in the evaluation (see below), it is of theoretical interest because its numerical similarity values can be interpreted in terms of evidence and uncertainty.

First of all, we assume that each model element represents a certain biochemical entity ("intended concept"), which the curator had in mind when assigning the annotations and which is described, possibly roughly or even wrongly, by the annotations. Furthermore, we assume that two intended concepts are classified as "alike" if they are biochemically close enough to be matched in model merging (the exact criterion for alikeness does not play a role for our argument). The annotations of two model elements can help to decide whether the two corresponding intended concepts are alike or not.

Our similarity score is supposed to quantify the *probability* that the intended concepts are alike, taking into account their annotations. If there is strong evidence that the intended concepts are alike, the model elements will obtain a similarity score close to 1. If there is little such evidence – or even evidence that the entities are not alike – the similarity score will be close to 0. To derive a formula for this probability, we first consider the simple case of two elements, each with a single annotation, and a single relation path between the corresponding BCs[2]. If there is no sufficiently short ontology path between the two BCs, we replace the path by a hypothetical "is_not" relation. Furthermore, if both elements refer to the same BC, the path consists of the self-relation "is". The three pieces of information – the two qualifiers and the BC relation – form a *semantic chain* $\lambda(\mu^{\mathtt{Q}}, \mathtt{path}(\mu, \nu), \nu^{\mathtt{Q}})$ between two elements. Some chains will provide strong evidence for alikeness (e.g., the triple "is/is/is"), while others provide little evidence (e.g., "is version/is/has part") or even evidence for non-alikeness ("is/is_not/is").

This evidence can be quantified by probabilistic formulae. We assume a hypothetical, large collection of annotated model elements. Pairs of elements are categorised as "alike" (A) or "not alike" (B) according to the intended concepts they represent. As pointed out

[2] Note that the BCs referenced by the annotations need not be identical with the intended concepts. For instance, they can be less specific than the intended concept because the curator could not find a web resource entry that would exactly match her intention.

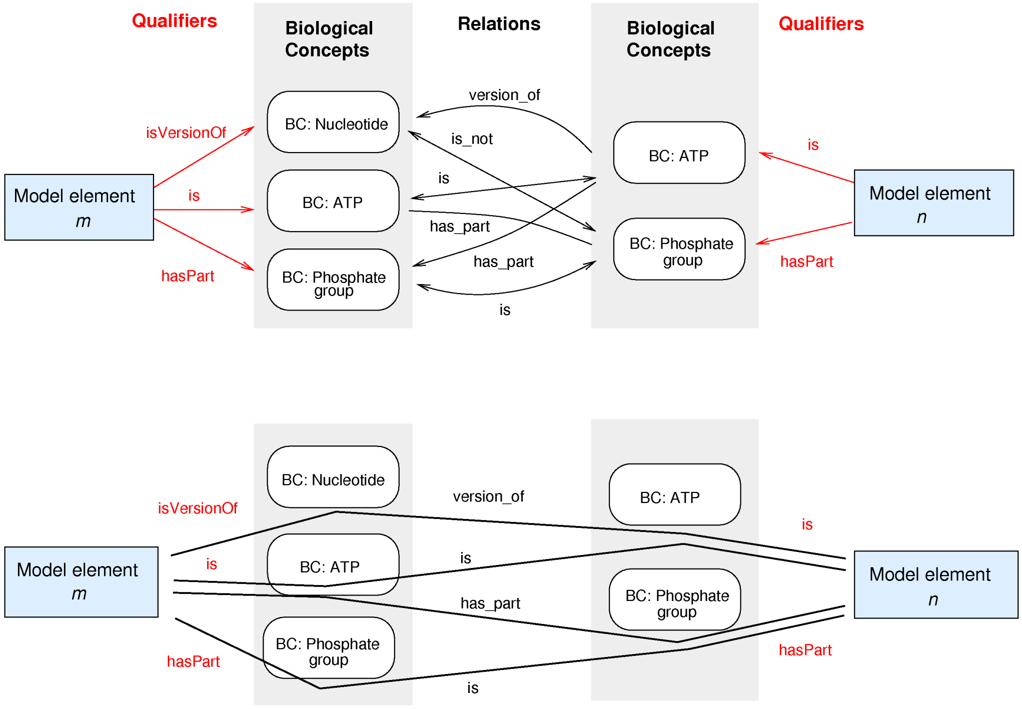


Figure 3: Semantic chains between two model elements. Top: Model elements (left and right boxes) refer to Biological Concepts (centre boxes) by biological qualifiers (red arrows). The Biological Concepts are interlinked via direct relations stored in the libSBAnnotation ontology and indirect relations derived from them (black arrows). Bottom: semantic chains between the two model elements, consisting of a qualifier, a relation, and another qualifier. Four of the possible chains are shown.

before, the exact criterion for alikeness does not play a role for our argument. If we consider a particular semantic chain λ (e.g., “is version/is/is version”), we can count how often this chain appears (or does not appear) in element pairs of type A or B. The probability to find the chain λ in a randomly picked pair of alike entities is given by the conditional probability $p(\lambda|A)$. The resulting log odds ratio

$$\tau_{\mathrm{El}}(\lambda) = \log \frac{p(\lambda|A)}{p(\lambda|B)} \tag{13}$$

tells us how much information about the question “alike or not alike?” is provided by the chain λ. If a chain is equiprobable for pairs of type A or B, the log odds ratio will be zero. If a chain appears more frequently in pairs of type A (B), the ratio will be positive (negative).

The next step is to combine information from several chains that might exist between the two model elements. This is illustrated in Figure 3. If each element contains several annotations and if pairs of referenced BCs are linked by several semantic chains, each possible path between the two model elements will contribute some information about the question “alike or not alike?”. The evidence from all these semantic chains can be combined by using the Bayesian formula. To determine the probability that the two elements are alike, we enumerate all combinatorial chains $\lambda_1, \lambda_2, ...$ and compute the posterior log odds ratio

$$L(m,n) \quad = \quad \log \frac{p(A|\lambda_1, \lambda_2, ...)}{p(B|\lambda_1, \lambda_2, ...)} = \log \frac{p(\lambda_1, \lambda_2, ...|A)}{p(\lambda_1, \lambda_2, ...|B)} + \log \frac{p(A)}{p(B)}. \tag{14}$$

As an approximation, we assume that the conditional probabilities of the chains are independent of each other and obtain

$$\begin{aligned} L(m,n) \approx L'(m,n) \quad &= \quad \log \frac{p(\lambda_1|A)}{p(\lambda_1|B)} + \log \frac{p(\lambda_2|A)}{p(\lambda_2|B)} + \cdots + \log \frac{p(A)}{p(B)} \\ &= \quad \tau_{\mathrm{El}}(\lambda_1) + \tau_{\mathrm{El}}(\lambda_2) + ... + \tau_{\mathrm{El}}. \end{aligned} \tag{15}$$

This formula consists of a sum of likelihood log odds ratios $\tau_{\mathrm{El}}(\lambda_i)$ for each chain plus a prior log odds ratio τ_{El} for two arbitrary elements to be alike. The values of all these log odds ratios can be collected in a table (for an example, see Table 5). Positive values will increase the probability for two elements to be alike, while negative ones will decrease it. Most chains can be assumed to have negligible values and are therefore omitted from the sum in Equation (15). If we accept the approximation (15) and translate the log ratios back to probabilities, we obtain our probabilistic similarity measure

$$\sigma_{\mathrm{El}}^{\mathrm{Prob}}(m,n) \quad = \quad \frac{e^{L'(m,n)}}{1 + e^{L'(m,n)}} \approx \mathrm{prob}(A|\lambda_1, \lambda_2, ...) \tag{16}$$

for model elements. Since we distinguished here between the *intended concepts* and the BCs representing them, we can also account for misannotations, where BCs were just wrongly assigned. A certain rate of such misannotations can be considered by simply changing the values of the conditional probabilities $p(\lambda_i|A)$.

5 Calculation of p-values for model similarities

5.1 Background model

A model similarity search yields a list of retrieved models, ranked by their similarities to the query model. Similarity values close to zero indicate that two models are basically dissimilar,

$\mu^{\mathtt{Q}}$	$\mathtt{path}(\mu,\nu)$	$\nu^{\mathtt{Q}}$	$\tau_{\mathrm{El}}(\lambda)$
is	has_functional_parent	is	1.
is	is	is	5.
is	is	isPartOf	1.
is	is	isVersionOf	2.
is	is_a	isVersionOf	1.
is	is	isHomologTo	2.
hasPart	is	hasVersion	0.5
hasPart	is	is	1.
hasPart	is_a	is	0.5
hasPart	is	isPartOf	1.3
hasPart	is	isVersionOf	0.5
hasPart	is_a	isVersionOf	0.2
hasPart	is	hasPart	1.3
hasPart	is_a	hasPart	0.6
hasPart	is	isHomologTo	1.
hasVersion	is	is	2.
hasVersion	is	hasVersion	1.
hasVersion	is	isVersionOf	1.
hasVersion	is	isHomologTo	1.
isDescribedBy	is	isDescribedBy	5.
isHomologTo	is	isHomologTo	2.5
isHomologTo	is	isPartOf	1.
isHomologTo	is	isVersionOf	1.5
isPartOf	is	isVersionOf	0.7
isVersionOf	is	isVersionOf	1.
isVersionOf	is_a	isVersionOf	0.5
	others		0

Table 5: Scores for semantic chains used in the probabilistic similarity measure $\sigma_{\mathrm{El}}^{\mathrm{Prob}}$. A chain $\lambda = (\mu^{\mathtt{Q}}, \mathtt{path}(\mu,\nu), \nu^{\mathtt{Q}})$ between both MIRIAM annotations μ^{A} and ν^{A} consists of the two qualifiers $\mu^{\mathtt{Q}}$ and $\nu^{\mathtt{Q}}$ and the relation between the referenced Biological Concepts μ and ν. In this example table, only direct relation arrows are listed. Semantic chains that are not listed in the table are assigned values of 0. For the prior log odds ratio $\tau_{\mathrm{El}} = \log \frac{p(A)}{p(B)}$, we choose a value of -3. All numerical values were chosen ad-hoc.

except for some elements that they possibly share by chance. But where can we draw the line between "similar" and "dissimilar" models? A pragmatic solution is to cut the list of retrieved models below a certain threshold value, e.g. a minimal similarity score of 0.25, chosen ad-hoc or from experience. While this might work well in practise, the choice of the threshold is arbitrary and a single threshold may not be appropriate for models of different size.

Alternatively, we can judge each similarity score by a p-value, describing the probability that the observed score could have occurred just by chance. To define precisely what is meant by "chance", we consider the similarities between our query model and an ensemble of random models. For the vector-based similarity measures, the random models can be simply defined by their feature vectors. For each possible BC, we count how many models in BioModels Database (x_i, total model number x_{tot}) refer to it in their annotations. From the count number, we compute its appearance probability

$$p_i = \mathrm{Prob}(v_i = 1) = \frac{x_i + 1}{x_{\mathrm{tot}} + 1} \tag{17}$$

where we use pseudo counts to avoid zero probability values. A random model is described by an independent random vector v in which each component v_i is set to a value $v_i = 1$ with probability p_i and to $v_i = 0$ otherwise. For a given query model, a background distribution for a similarity score $\sigma_{\mathrm{Mo}}^{\mathrm{TVSM}}$ (as shown in Figure 4) is derived by evaluating the scores between the query and the random models.

The proposed null model was specifically chosen to emphasise the "unexpected" common appearance of BCs shared by the query model and the retrieved models. By construction, a low p-value (e.g. below 0.05) means that the observed similarity score cannot simply be explained by random appearance of individual BCs with their different frequencies. Instead, it hints at a correlated appearance of BCs in both models, which suggests that the two models describe similar biological pathways. Furthermore, using this background model we take care of the fact that models could overlap in common molecules like nucleotides by requiring a higher similarity threshold e.g. for models in which ATP appears.

If we score the retrieved models by their p-values (instead of the similarities), the frequencies of individual BCs are, as mentioned above, taken into account in the background model. Therefore, the term f_3 in the formula for BC similarities would provide no additional information and could be omitted.

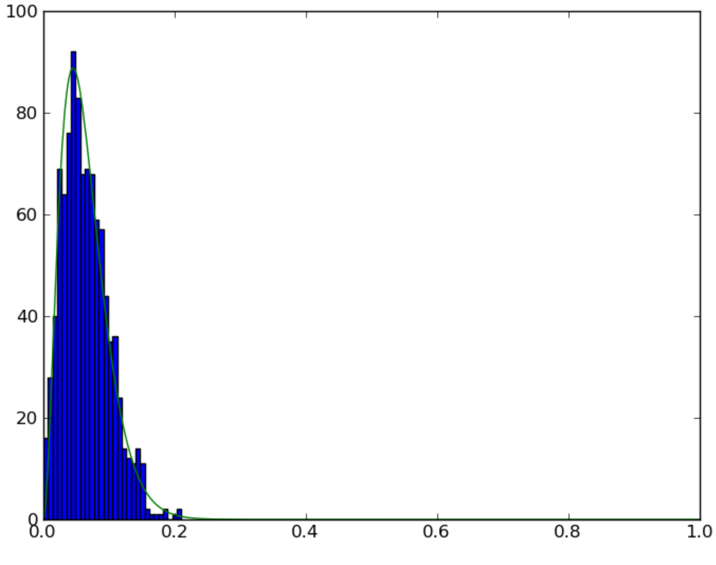


Figure 4: Calculation of p-values for model similarities. The blue histogram shows the distribution of vector-based similarity scores $\sigma_{\mathrm{Mo}}^{\mathrm{TVSM}}$ between the BioModel 9 [3] and 1000 randomly generated feature vectors representing random models. The distribution resembles a beta distribution (green, fitted) and is concentrated at values below 0.2. Given these random similarities the observed ones of the first retrieved models (Figure 3 in main article) are highly significant.

5.2 Bayesian estimation of p-values

Even for this simple ensemble of random models, the background distribution for our similarity measure is hard to compute as there exists no closed, simple formula. Only with a few simplifications concerning the similarity measure and the feature vectors an efficient explicit computation becomes possible.

However, p-values for any similarity measure can be estimated by sampling many realisations of the random model and employing a Bayesian approach. The number x_{sim} counts the positive random models, i.e. the models showing a higher similarity than the retrieved model. Under the null hypothesis, the number x_{sim} of positive random models would follow a binomial distribution with distribution parameter p and maximal number x_{rand}. Given x_{sim} and x_{rand} and assuming a uniform prior distribution for the p-value P, the posterior for P is a beta distribution $\text{prob}(P) \sim P^{x_{\text{sim}}}(1-P)^{x_{\text{rand}}-x_{\text{sim}}}$, which describes our estimated p-value. Its mean value and standard deviation read

$$\begin{aligned}\langle P \rangle &= \frac{x_{\text{sim}}+1}{x_{\text{rand}}+2}\\ \sqrt{\text{var}(P)\rangle} &= \sqrt{\frac{(x_{\text{sim}}+1)(x_{\text{rand}}-x_{\text{sim}}+1)}{(x_{\text{rand}}+2)^2(x_{\text{rand}}+1)}} \approx \frac{\sqrt{x_{\text{sim}}+1}}{x_{\text{rand}}+2} \qquad (18)\end{aligned}$$

Thus, the estimation of low p-values is limited by the number x_{rand} of random models considered: for $x_{\text{rand}} = 998$ even a similarity of 1 would be assigned a p-value of 0.001 ± 0.001.

5.3 Analytic derivation of p-values

There is no simple analytic formula for the p-value of our vector based similarity measure. This is due to the mathematical structure of the similarity formula and the fact that all BCs can have different occurrence probabilities in a random model. In order to deal with the first problem, we start to develop a way to compute p-values for a more simple similarity measure and show ways to extend it, while we tackle the second problem by an efficient computation.

p-value for model overlap First we consider the simple model overlap $\sigma^{\text{O}}_{\text{Mo}}(M,N) = v_M^{\text{T}} v_N$, which counts the number of BCs which are referred to in both models under the condition that the vectors contain only zeros and ones. Without loss of generality, let the BCs be sorted such that $\forall_{i \in 1..|M|} v_{i\text{M}} = 1$ and $v_{i\text{M}} = 0$ otherwise. With p_i being the probability of an annotation to occur in a model we can easily determine the probability of a random model N to have an overlap of zero to model M:

$$\Pr\big(\sigma^{\text{O}}_{\text{Mo}}(M,N) = 0\big) = \prod_{i=1}^{|M|} (1-p_i)$$

or the maximal overlap of $|M|$:

$$\Pr\big(\sigma^{\text{O}}_{\text{Mo}}(M,N) = |M|\big) = \prod_{i=1}^{|M|} p_i.$$

The probability distribution for random models referring to a certain number of the first $|M|$ BCs can be obtained by convolution of Bernoulli distributions describing the probabilities of containing the single BCs. Computationally this can easily be solved by dynamic programming.

The dynamic programming matrix D is filled with the conditional probabilities of a random model referring to a certain number of the first $|M|$ BCs given $|M|$.

$$D_{0,0} = \Pr\bigl(\sigma^{\mathrm{O}}_{\mathrm{Mo}} = 0 \bigm| |M| = 0\bigr) = 1$$

is the trivial anchor of the iteration, and

$$\begin{aligned} D_{x,y} &= \Pr\bigl(\sigma^{\mathrm{O}}_{\mathrm{Mo}} = x \bigm| |M| = y\bigr) \\ &= p_y \cdot \Pr\bigl(\sigma^{\mathrm{O}}_{\mathrm{Mo}} = x-1 \bigm| |M| = y-1\bigr) + (1-p_y) \cdot \Pr\bigl(\sigma^{\mathrm{O}}_{\mathrm{Mo}} = x \bigm| |M| = y-1\bigr) \\ &= p_y \cdot D_{x-1,y-1} + (1-p_y) \cdot D_{x,y-1} \end{aligned}$$

defines the stepwise completion of D. Since the final p-value is given by $1 - \sum_{i=0}^{\sigma^{\mathrm{O}}_{\mathrm{Mo}}(M,N)-1} D_{i,|M|}$, the matrix D does not need to be computed completely. Therefore, the computational effort of the calculation is relatively low ($\mathcal{O}(|M| \cdot \sigma^{\mathrm{O}}_{\mathrm{Mo}}(M,N))$). In case many BCs are associated identical p_i values, the computation can be sped up by convoluting binomial instead of Bernoulli distributions.

p-value for querying the database The above mentioned p-value describes the probability of observing a certain or higher score by chance in a random model. In case we would not be interested in individual models but rather would like to know whether significant parts of a certain model are already described by available models, we could extend this to a p-value describing the probability of finding at least one model with this or a higher score in the database. The formula for this extended p-value reads

$$p_{\mathrm{e}} = 1 - (1-p)^{|\mathcal{M}|},$$

where $|\mathcal{M}|$ is the number of models in the database and p is the p-value for the comparison to one single random model.

Accounting for similar annotations In the overlap measure $\sigma^{\mathrm{O}}_{\mathrm{Mo}}$ we disregard the knowledge about similar BCs. The matrix S, containing this information, introduces a new level of complexity to the p-value calculation for a similarity measure $\sigma^{\mathrm{OS}}_{\mathrm{Mo}}(M,N) = v_{\mathrm{M}}^{\mathrm{T}} S v_{\mathrm{N}}$, because real valued similarity scores become possible. Including probabilities of these real valued similarities would blow up the size of the dynamic programming matrix. A pragmatic solution to this problem is to allow only a limited number of rational numbers, e.g. fractions of 10, in the S matrix, building a bigger D matrix accounting for the new similarity scores, and customising the iteration step integrating the score contribution of every single BC.

Accounting for vector lengths Our complete similarity measure $\sigma^{\mathrm{TVSM}}_{\mathrm{Mo}}$ also accounts for the length of the vectors v_{M} and v_{N} which has previously not be considered. In order to incorporate this normalisation, we now have to consider two random variables describing the number of BCs in model N which also appear in M and those who do not:

$$X = \sum_{i=1}^{|M|} v_{i\mathrm{N}}$$

$$Y = \sum_{i=|M|+1}^{|A|} v_{i\mathrm{N}},$$

where $|A|$ is the length of the feature vectors, i.e., the number of all BCs which are mentioned in any model of the BioModels Database +1 (representing BCs not referred to). The probability distribution of these variables can be calculated as explained above, although the complete dynamic programming matrices have to be constructed. Since both variables are independent, their combined probabilities can be constructed by multiplying their single probabilities. Finally, the probabilities of those X and Y pairs for which $\frac{X}{\sqrt{X+Y}} \geq \sqrt{|M|}\sigma_{\mathrm{Mo}}^{\mathrm{TVSM}}(M,N)$ are summed up to yield the p-value.

With the extensions explained here the computation of p-values for our similarity measure is in principle possible. Nevertheless, the computation is too slow for this method to be included into our web service and thus only the explicitly computed p-value for the model overlap is shown. In contrast to the Bayesian way of computing p-values, this explicit approach can only be used for simple similarity measures. More complex similarity measures as the structured ones introduced in this article or even network structure based ones might also require a different, more complex null model. As they include more or different kinds of information a bias in the similarity measure towards certain models will arise. Without accounting for this bias in the null model and therefore in the computation of the p-value such a similarity measure will lose its value as the number of available models grows.

6 Evaluation of similarity measures

Similarity measures do not neutrally describe things “as they are”, but emphasise our human perspective and are chosen to serve certain purposes. In our case, similarities will be useful if they match the judgement of human experts whether two models describe similar or closely related biological processes or pathways. To evaluate this for our similarity measures, we collected benchmark models, classified them into predefined biological groups, and considered model clustering as a potentially important test case.

6.1 Benchmark: model sets with predefined biological groups

As a gold standard for statistical evaluation, we chose two sets of benchmark models and grouped them by biological categories.

1. **Small benchmark set.** We considered 14 manually selected models representing four distinct types of biochemical systems: glycolysis (models 70, 71, 211), circadian clock (models 16, 21, 22), cell cycle (models 5, 7, 8, 111), and MAP kinase pathways (models 26, 27, 28, 29).

2. **Large benchmark set.** For more comprehensive tests, we considered the entire BioModels Database (16$^{\mathrm{th}}$ release). Since this model set was too big to predefine model groups by hand, we grouped them semi-automatically according to the MIRIAM annotations of their SBML `<model>` elements (see Table 6). Of course, these model annotations are not taken into account when computing the model similarities.

To test our similarity measures, we computed the pairwise similarities between all models in the benchmark sets, used them for unsupervised clustering, and compared all results to the predefined model groups.

6.2 Quality criteria used for evaluation

Model similarity: evaluation by silhouette coefficient Models from one biological group should be more similar than models from different groups. To evaluate our measures, we determined the average intra- and inter-group similarities and scored them by the silhouette coefficient [6]

$$\mathtt{sc}(\mathcal{M}) = \frac{\sum_{M \in \mathcal{M}} \frac{\iota(M) - \epsilon(M)}{\max(\iota(M), \epsilon(M))}}{|\mathcal{M}|}, \tag{19}$$

where

$$\begin{aligned}
\epsilon(M) &= \max_{\mathcal{C} \in \mathbf{C}, M \notin \mathcal{C}} \frac{\sum_{N \in \mathcal{C}} \sigma_{\mathrm{Mo}}(M, N)}{|\mathcal{C}|}, \\
\iota(M) &= \max_{\mathcal{C} \in \mathbf{C}, M \in \mathcal{C}} \frac{\sum_{N \in \mathcal{C}, N \neq M} \sigma_{\mathrm{Mo}}(M, N)}{|\mathcal{C}|},
\end{aligned} \tag{20}$$

$\mathcal{M}$ is the set of benchmark models, and $\mathbf{C}$ is the set of predefined biological model groups. The silhouette coefficient compares the similarities of models within (ι) and between (ϵ) groups and indicates how clearly models from different groups are separated by the similarity measures.

Model clustering: evaluation by Jaccard coefficient A clustering based on model similarities should split the models into biologically meaningful groups. As a test, we clustered the benchmark models by an agglomerative clustering with average linkage, based on the different model similarity measures, and cut the dendrogram at a certain height. The cut height was chosen such that the numbers of clusters and predefined groups were identical. We compared the resulting clusters with the biological groups by the Jaccard similarity coefficient [4]

$$\mathtt{jac} = \frac{O_{11}}{O_{01} + O_{10} + O_{11}}. \tag{21}$$

In this formula, O_{11} is the number of model pairs that share the same group and the same cluster, while O_{10} and O_{01} count the pairs that share only the same group or the same cluster, respectively.

6.3 Systematic evaluation of the similarity measures

We considered different variants of our similarity measures (all normalised and with the numerical parameters listed in Tables 2, 3, 4, and 5) and first evaluated them with the small benchmark set. In general, all measures performed well and the differences between them were rather small. Both the similarity matrices (Figure 5) and the clustering results (Figure 6) show a good agreement with the predefined biological groups: model similarity within groups tends to be high, while models from different groups show little similarity. The cell cycle models, however, appeared to be rather dissimilar and shared some annotations with the circadian clock models. Some minor differences between the four similarity measures are visible in Figure 5. While the results for Li's measure $\sigma_{\mathrm{BC}}^{\mathrm{Li}}$ and the distance-depth-dependent measure $\sigma_{\mathrm{BC}}^{\mathrm{DD}}$ look very similar, the preference-based measure based on probabilistic element similarities $\sigma_{\mathrm{El}}^{\mathrm{Prob}}$ shows a stronger background and a weaker intra-group similarity, but also the inter-group similarity is less prominent. On the contrary, the vector-based measure $\sigma_{\mathrm{Mo}}^{\mathrm{TVSM}}$ shows a stronger intra-group similarity, but also a high inter-group similarity across the cell cycle, circadian clock, and MAPK models.

For a more systematic evaluation, we considered the large benchmark set, calculated all pairwise similarities, and evaluated them with the silhouette (Eq. 19) and Jaccard (Eq. 21)

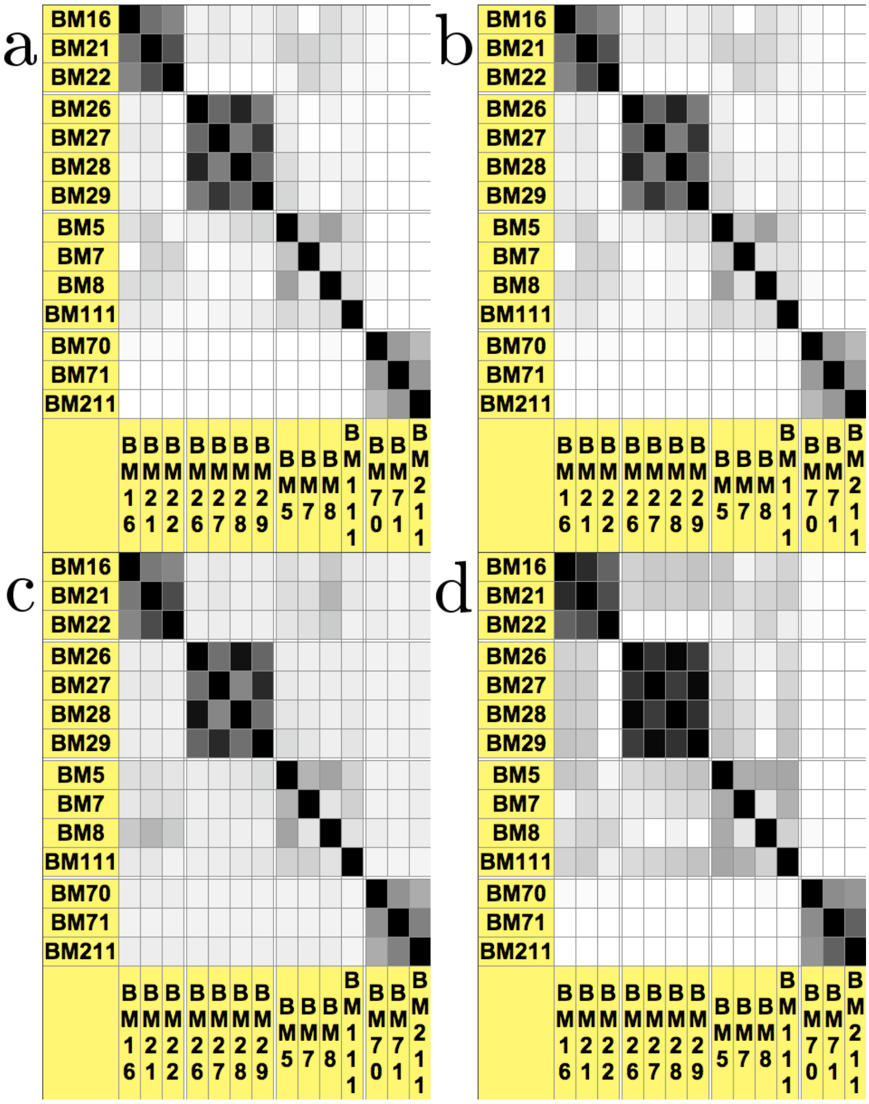
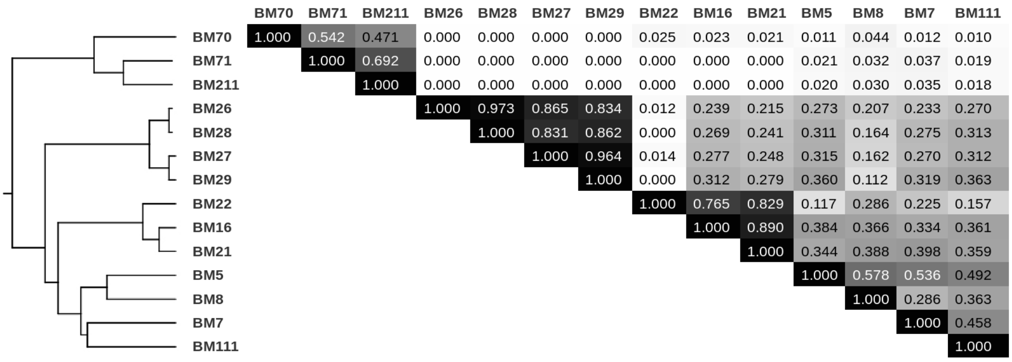


Figure 5: Similarity matrices for the small set of benchmark models. Model similarities were computed by the preference-based similarity measure $\sigma_{\mathrm{Mo}}^{\mathrm{Pref}}$ (a) with independent element similarity $\sigma_{\mathrm{BC}}^{\mathrm{Li}}$, (b) distance-depth-dependent element similarity $\sigma_{\mathrm{BC}}^{\mathrm{DD}}$, and (c) probabilistic similarity $\sigma_{\mathrm{El}}^{\mathrm{Prob}}$. The vector-based measure $\sigma_{\mathrm{Mo}}^{\mathrm{TVSM}}$ without semantic density term (i.e. $f_3 = 1$) is shown in (d). Colours indicate similarity values from 0 (white) to 1 (black).



	BM70	BM71	BM211	BM26	BM28	BM27	BM29	BM22	BM16	BM21	BM5	BM8	BM7	BM111
BM70	1.000	0.542	0.471	0.000	0.000	0.000	0.000	0.025	0.023	0.021	0.011	0.044	0.012	0.010
BM71		1.000	0.692	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.021	0.032	0.037	0.019
BM211			1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.020	0.030	0.035	0.018
BM26				1.000	0.973	0.865	0.834	0.012	0.239	0.215	0.273	0.207	0.233	0.270
BM28					1.000	0.831	0.862	0.000	0.269	0.241	0.311	0.164	0.275	0.313
BM27						1.000	0.964	0.014	0.277	0.248	0.315	0.162	0.270	0.312
BM29							1.000	0.000	0.312	0.279	0.360	0.112	0.319	0.363
BM22								1.000	0.765	0.829	0.117	0.286	0.225	0.157
BM16									1.000	0.890	0.384	0.366	0.334	0.361
BM21										1.000	0.344	0.388	0.398	0.359
BM5											1.000	0.578	0.536	0.492
BM8												1.000	0.286	0.363
BM7													1.000	0.458
BM111														1.000

Figure 6: Cluster tree of the small benchmark set. Hierarchical clustering with average linkage based on the vector-based similarity measure $\sigma_{\mathrm{Mo}}^{\mathrm{TVSM}}$ without semantic density (i.e., setting $f_3 = 1$). The predefined model groups are well matched by the cluster tree obtained from the model similarities (compare Figure 5 (d)).

coefficients. The tests revealed differences, but not a clear ordering in their quality. As shown in Table 7, the quality of the similarity measures depends on whether they are judged by the silhouette or the Jaccard coefficient and whether we consider the small or large model benchmark set.

In the large benchmark set, the vector-based measure without semantic density performed better both for the silhouette and the Jaccard coefficients. While the independent and the distance-depth-dependent measure perform similarly, the original Li measure performs better when the local semantic density (in the term f_3) is included. This fact might be incidental as the results from the other measures suggest that the semantic density does not improve the quality of the measure. Among the vector-based measures, it seems that the inclusion of the term f_3 impairs the results. This finding was unexpected, but agrees with the observations of Li *et al.* for similarity of phrases [7]. Furthermore, the probabilistic preference-based measure performs weakest in this comparison, both for the Jaccard coefficient of the large benchmark set and for the silhouette coefficient of the small benchmark set.

Surprisingly, the silhouette coefficient partially increases between the cases $S = f_1^{f_2}$ and $S = I$ in Table 7, implying that model groups cannot necessarily be better distinguished if all semantic knowledge from the libSBAnnotation is considered. A possible reason is that the annotations in BioModels database are very consistent between models because these models are annotated by a small number of professional curators. Accordingly, a matching of slightly different annotations will be less important than for models stemming from different sources and containing less consistent annotations. Altogether, proving or disproving the usefulness of semantic ontology information for model classification would probably require larger consistent benchmark sets with less consistent semantic annotations.

The vector-based similarity scores, which are simple and easy to compute, performed well in the comparison. However, also the preference-based measures have their advantages as they make use of the information to which model elements annotations are assigned to. Furthermore, all measures can be customised by changing their parameter values. In particular, parameters could be optimised, e.g., by comparing clustering results to the our gold standard classification by the Jaccard coefficient and improving this value by changes in the parameters. We tried to do this, but the results appeared to be overfitted. The current

<model> annotation	**Name**	**Models**
GO:0019228	regulation of action potential in neuron	124, 127, 129, 130, 131, 132, 133, 134, 135, 136, 141, 142
GO:0006096 kegg.pathway:sce00010	glycolysis	42, 51, 61, 63, 64, 70, 71, 172, 176, 177, 206, 211, 225
GO:0048863	stem cell differentiation	203, 204, 209, 210
GO:0006915 kegg.pathway:hsa04210	apoptosis	102, 103, 220
GO:0005248 GO:0019227 GO:0005249	voltage-gated sodium channel activity, neuronal action potential propagation, voltage-gated potassium channel activity	20, 118, 119
GO:0048511	rhythmic process	79, 99
GO:0009755 GO:0051924	regulation of calcium ion transport, hormone-mediated signalling	114, 115
GO:0007259 kegg.pathway:mmu04630	JAK-STAT cascade	93, 94, 151
GO:0019236	response to pheromone	32, 116
reactome:REACT_634	MAP kinase cascade	9, 10, 11, 14
GO:0016692	NADH peroxidase activity	46, 143
GO:0007188	G-protein signalling, coupled to cAMP nucleotide second messenger	128, 165
GO:0045990	regulation of transcription by carbon catabolites	65, 67
kegg.pathway:mmu04660	T cell receptor signalling	139, 140, 147, 226, 227, 230
GO:0009088 kegg.pathway:map00260	threonine biosynthetic process	66, 68
GO:0008277	regulation of G-protein coupled receptor protein signalling pathway	85, 86
GO:0006099 kegg.pathway:ko00020 GO:0006097 reactome:REACT_1785	glyoxylate cycle, tricarboxylic acid cycle	218, 219, 222
GO:0019722 kegg.pathway:hsa04020 kegg.pathway:map04020	calcium-mediated signalling	39, 43, 44, 45, 47, 57, 58, 59, 60, 81, 100, 113, 117, 145, 166, 184
GO:0031684	heterotrimeric G-protein complex cycle	72, 80, 82
GO:0006935	chemotaxis	200, 229
kegg.pathway:hsa04012	ErbB signalling pathway	175, 223
GO:0006816	calcium ion transport	98, 162
GO:0000278 kegg.pathway:sce04111 kegg.pathway:hsa04110 reactome:REACT_152	mitotic cell cycle	3, 4, 5, 6, 7, 8, 56, 69, 87, 107, 109, 110, 111, 144, 150, 168, 181, 186, 187, 193, 194, 196, 207, 208
kegg.pathway:hsa04660	T cell receptor signalling	120, 122, 123
GO:0007623 kegg.pathway:hsa04710	circadian rhythm	16, 21, 22, 24, 25, 34, 36, 55, 73, 74, 78, 83, 89, 95, 96, 97, 160, 170, 171, 214, 216
GO:0016055	Wnt receptor signalling pathway	149, 201
GO:0007173	epidermal growth factor receptor signalling pathway	19, 33, 48, 49, 84, 161
kegg.pathway:hsa04115	p53 signalling pathway	154, 155, 156, 157, 158, 159, 188, 189
GO:0007166	cell surface receptor linked signal transduction	1, 2, 125
GO:0046655	folic acid metabolic process	18, 213
GO:0002028	regulation of sodium ion transport	54, 126
kegg.pathway:hsa04350	TGF-beta signalling pathway	101, 112, 163, 173
GO:0040029	regulation of gene expression, epigenetic	12, 104
from small example	MAPKKK cascade	26, 27, 28, 29

Table 6: Large set of benchmark models. Models from BioModels Database were semi-automatically classified into joint biological groups taking into account the MIRIAM annotations of their <model> elements. Some annotations, e.g. GO:0000165 (MAPKKK cascade) or the annotations for organisms, referring to the NCBI Taxonomy, would have resulted in too big clusters and were therefore ignored.

BioModels database seems to be too small and restricted in the types of models included. As long as good training data for comparison is lacking, the estimation of all parameters, especially for the dependent preference-based method, is not possible.

Similarity measure $\hat{\sigma}$	**Silhouette coefficient**		**Jaccard coefficient**	
(normalised)	**Small set**	**Large set**	**Small set**	**Large set**
TVSM, S given by $\sigma_{\mathrm{BC}}^{\mathrm{DD}}$	.657	.0982	1	.284
with $S = f_1^{f_2}$	.705	.136	1	.377
with $S = I$	.766	.146	1	.356
Preference-based with $\sigma_{\mathrm{El}}^{\mathrm{Pref}}$ and $\sigma_{\mathrm{BC}}^{\mathrm{Li}}$	.746	.108	1	.231
setting $c_\mu^c = c_\mu$	.746	.108	1	.231
setting $f_3^{\mathrm{Li}} = 1$	.746	.104	1	.229
Preference-based with $\sigma_{\mathrm{El}}^{\mathrm{Pref}}$ and $\sigma_{\mathrm{BC}}^{\mathrm{DD}}$	.741	.118	1	.231
setting $c_\mu^c = c_\mu$	.738	.120	1	.231
setting $f_3 = 1$	.700	.123	1	.269
additionally setting $f_2 = 1$	.720	.122	1	.254
additionally without libSBAnnotation	.746	.101	1	.229
additionally setting $f_{\mathrm{qsm}} = 1$	.709	.095	.556	.262
Preference-based, using $\sigma_{\mathrm{El}}^{\mathrm{Prob}}$	.663	.141	1	.207

Table 7: Evaluation of model similarity measures with predefined model groups. Different variants of normalised similarity measures (rows) were compared for the small and large model benchmark sets. The silhouette coefficient scores the similarities of models within and between groups. For computing the Jaccard coefficient, the models were clustered by agglomerative clustering with average linkage and with the respective similarity measure. The dendrograms were cut at a height where the numbers of clusters and predefined model groups were identical (4 groups for the small benchmark set; 34 for the large benchmark set).

6.4 Ranking result depend little on the values of relation type scores

Although we have been unable to meaningfully optimise the parameters used in our similarity measures we have performed a sensitivity analysis for the retrieval results described in Figure 3 in the main text with respect to the relation type scores f_{rts}. For this purpose we randomly constructed 100 sets of relation type scores by multiplying each of the entries with a random normally distributed variable centred around 1 (resulting f_{rts} values were restricted to the range $[0, 1]$) and started a model retrieval for models similar to BioModel 9. Across the 100 results we determined mean and standard deviation for the similarity score and the rank of the retrieved model.

The results for this example (see Table 8) suggest that the ranking and the similarities are only slightly influenced by the f_{fts} values. This effect might be due to the above mentioned fact that all models have been annotated by the same curators and in most cases use not only similar, but identical annotations.

7 Application: Which part of metabolism is covered by kinetic models?

An ambitious aim of Systems Biology is to construct large-scale dynamic models of cellular metabolism. In such models, the metabolic network would be populated with quantitative formulae for the enzymatic rate laws. Herrgård *et al.* [2] have compiled a metabolic net-

Model	SD = .1				SD = .5			
	Rank		Similarity score		Rank		Similarity score	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
9	1	0	1	0	1	0	1	0
11	2	0	.925	.0003	2	0	.925	.0022
14	3	0	.865	.0006	3	0	.865	.0037
10	4	0	.816	.0009	4	0	.816	.0058
26	5	0	.737	.0013	5	0	.736	.0087
28	6	0	.687	.0025	6.08	.392	.685	.0174
30	7	0	.687	.0025	7.08	.392	.685	.0174
27	8	0	.673	.0018	7.92	.392	.672	.0122
31	9	0	.673	.0018	8.92	.392	.672	.0122
29	10	0	.614	.0033	10	0	.612	.0228
84	11	0	.482	.0049	11	0	.480	.0291
116	12	0	.397	.0029	12	0	.396	.0177
32	13	0	.348	.0028	13	0	.346	.0182
149	14	0	.335	.0023	14.09	.286	.333	.0158
205	15	0	.299	.0079	15.24	.991	.298	.0455
33	16	0	.260	.0010	15.84	.367	.259	.0065
16	17	0	.244	.0027	17.16	.367	.242	.0171
49	18	0	.240	.0015	17.75	.639	.239	.0098
21	19	0	.230	.0025	19.02	.316	.228	.0161
4	20	0	.222	.0031	20.16	.463	.220	.0199

Table 8: Sensitivity of similarities and model retrieval ranking with respect to relation type scores. Shown are the mean and the standard deviation of similarity and rank for the retrieved models when searching for models similar to BioModel 9. Mean and standard deviation are determined in 100 trials in which each relation type score is multiplied by a Gaussian distributed random variable with mean 1 and standard deviation (SD) 0.1 or 0.5.

work of the yeast *S. cerevisiae*, containing most of the known metabolic reactions in this common model organism. To study which fraction of this network can already be covered by existing kinetic models, we defined a preference-based overlap score $\omega^{\mathrm{Pref}}_{\mathrm{Mo}}(M,N) = \sigma^{\mathrm{Pref}}_{\mathrm{Mo}}(M,N) \cdot (|M|,|N|)$ between the yeast network and models in BioModels Database, counting approximately the number of shared model elements. Model similarities were computed based on the scores $\sigma^{\mathrm{Pref}}_{\mathrm{El}}$ and $\sigma^{\mathrm{DD}}_{\mathrm{BC}}$. As shown in Table 9, most of the high overlaps were achieved by models of central metabolism, in particular glycolysis. This is not surprising, since central metabolism has been a main field of biochemical research for many years. To cover further parts of the network, we chose the model with the highest overlap score (model number 239, [5]), extracted all its annotated elements and screened the yeast network for elements that had a similarity σ_{El} greater or equal to 0.3 to any of them. After masking these elements, the procedure was repeated to determine the next matching model. After eight iterations, the selected models covered about one seventh of the network, while all further models would contribute only few additional elements. As shown in Table 10, the selected models describe diverse metabolic pathways in different organisms. However, only two of them (BioModel 90 [13] and BioModel 172 [9]) actually represent metabolism in yeast[3].

Pathway described	**BioModels model number**	**Overlap score** $\omega^{\mathrm{Pref}}_{\mathrm{Mo}}$
Pancreatic beta cells	239	238
Glycolysis	172	113
Erythrocyte metabolism	70	107
Glycolysis	64	107
Glycolysis	177	106
Glycolysis	176	104
Respiratory oscillations	90	102
Glycolysis	61	101
Calvin cycle	13	85.7
Pyruvate branches	17	81.8
Aspartate metabolism	212	77.4
Folate cycle	18	65.4
Glycolysis	42	59.3
Glycolysis	71	57.4

Table 9: Models from BioModels Database showing a large overlap with the yeast consensus model [2]. The overlap score approximately describes the number of shared model elements (compounds and reactions).

Pathway described	**BioModels model number**	**New elements contributed**	**Elements covered**
Pancreatic beta cells	239	215	215
Respiratory oscillations	90	45	260
Aspartate metabolism	212	49	309
Erythrocyte metabolism	70	27	336
Folate cycle	18	37	373
Glycolysis	172	73	446
Purine metabolism	15	17	463
Polyamine metabolism	190	17	480

Table 10: Covering the yeast metabolic network with kinetic models. About one seventh of the yeast consensus model [2] (containing 3279 annotated elements) was successively covered by kinetic models selected from BioModels Database. While the first model shows an overlap of 215 network elements (compounds and reactions), the following models contribute less and less additional elements.

References

[1] J. Becker and D. Kuropka. Topic-based vector space model. In *Proceedings of the 6th International Conference on Business Information Systems*, pages 7–12, 2003.

[3] Currently, BioModels database contains 18 yeast models in total, covering mostly central metabolism and signalling pathways.

[2] M.J. Herrgård, N. Swainston, P. Dobson, W.B. Dunn, K.Y. Arga, M. Arvas, N. Büthgen, S. Borger, R. Costenoble, M. Heinemann, M. Hucka, N. Le Novère, P. Li, W. Liebermeister, M.L. Mo, A.P. Oliveira, D. Petranovic, S. Pettifer, E. Simeonidis, K. Smallbone, I. Spasić, D. Weichart, R. Brent, D.S. Broomhead, H.V. Westerhoff, B.I. Kirdar, M. Penttilä, E. Klipp, B.. Palsson, U. Sauer, S.G. Oliver, P. Mendes, J. Nielsen, and D.B. Kell. A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nature Biotechnology*, 26(10):1155–1160, 2008.

[3] CY Huang and J.E. Ferrell. Ultrasensitivity in the mitogen-activated protein kinase cascade. *Proceedings of the National Academy of Sciences*, 93(19):10078, 1996.

[4] P. Jaccard. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull. Soc. Vaudoise Sci. Nat*, 37:547–579, 1901.

[5] N. Jiang, R.D. Cox, and J.M. Hancock. A kinetic core model of the glucose-stimulated insulin secretion network of pancreatic β cells. *Mammalian Genome*, 18(6):508–520, 2007.

[6] L. Kaufman and PJ Rousseeuw. Finding groups in data; an introduction to cluster analysis. *Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics Section (EUA).*, 1990.

[7] Y. Li, Z.A. Bandar, and D. McLean. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on knowledge and data engineering*, pages 871–882, 2003.

[8] A.L. Lister, M. Pocock, M. Taschuk, and A. Wipat. Saint: a lightweight integration environment for model annotation. *Bioinformatics*, 25(22):3026, 2009.

[9] L. Pritchard and D.B. Kell. Schemes of flux control in a model of Saccharomyces cerevisiae glycolysis. *European journal of biochemistry*, 269(16):3894–3904, 2002.

[10] N. Swainston and P. Mendes. libAnnotationSBML: a library for exploiting SBML annotations. *Bioinformatics*, 25(17):2292, 2009.

[11] M.P. Van Iersel, A.R. Pico, T. Kelder, J. Gao, I. Ho, K. Hanspers, B.R. Conklin, and C.T. Evelo. The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. *BMC bioinformatics*, 11(1):5, 2010.

[12] C.J. van Rijsbergen. *Information Retrieval.* Butterworths, London, 2nd edition, 1979.

[13] J. Wolf, H.Y. Sohn, R. Heinrich, and H. Kuriyama. Mathematical analysis of a mechanism for autonomous metabolic oscillations in continuous culture of Saccharomyces cerevisiae. *FEBS letters*, 499(3):230–234, 2001.