

## PERSPECTIVE

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

Marvin Schulz<sup>1</sup>, Falko Krause<sup>1</sup>, Nicolas Le Novère<sup>2</sup>,  
Edda Klipp<sup>1,\*</sup> and Wolfram Liebermeister<sup>1,3</sup>

<sup>1</sup> Institut für Biologie, Theoretische Biophysik, Humboldt-Universität zu Berlin, Berlin, Germany,

<sup>2</sup> Computational Neurobiology, EMBL-EBI, Wellcome-Trust Genome Campus, Cambridge, UK and

<sup>3</sup> Department of Plant Sciences, Weizmann Institute of Science, Rehovot, Israel

\* Corresponding author. Institut für Biologie, Theoretische Biophysik, Humboldt-Universität zu Berlin, Invalidenstr. 42, Berlin 10115, Germany. Tel.: + 49 302 093 9040; Fax: + 49 302 093 8813; E-mail: edda.klipp@rz.hu-berlin.de

Received 21.9.10; accepted 31.5.11

**The exploding number of computational models produced by Systems Biologists over the last years is an invitation to structure and exploit this new wealth of information. Researchers would like to trace models relevant to specific scientific questions, to explore their biological content, to align and combine them, and to match them with experimental data. To automate these processes, it is essential to consider semantic annotations, which describe their biological meaning. As a prerequisite for a wide range of computational methods, we propose general and flexible similarity measures for Systems Biology models computed from semantic annotations. By using these measures and a large extensible ontology, we implement a platform that can retrieve, cluster, and align Systems Biology models and experimental data sets. At present, its major application is the search for relevant models in the BioModels Database, starting from initial models, data sets, or lists of biological concepts. Beyond similarity searches, the representation of models by semantic feature vectors may pave the way for visualisation, exploration, and statistical analysis of large collections of models and corresponding data.**

*Molecular Systems Biology* 7: 512; published online 19 July 2011; doi:10.1038/msb.2011.41

**Subject Categories:** metabolic and regulatory networks; computational methods

**Keywords:** BioModels database; ontology; semanticSBML; similarity measure

## Introduction

The rise of Systems Biology as a mainstream field of research triggered a fast accumulation of knowledge about cellular networks, their biochemical details, and their dynamic behaviour. Much of this complex information is condensed

in mathematical models, which statically or dynamically describe the interconversion of biochemical compounds within reaction networks. A wealth of models, picturing various regions of the cellular networks, are available in public repositories like the BioModels Database (Le Novère *et al*, 2006) or JWS Online (Olivier and Snoep, 2004) in the machine-readable format Systems Biology Markup Language (SBML; Hucka *et al*, 2003). Meta-information on existing databases can be found on websites like PathGuide (Bader *et al*, 2006). The models in these repositories serve as information sources and they may be reused, refined, and combined for new research studies.

Continued research aiming for improved and complex models, e.g., for biomedical purposes, makes it desirable to change or combine models automatically with the help of computers. Such an automatic processing would be easier if published models and data were based on a common list of well-defined elements with a fixed naming convention. However, since models cover a growing number of entities and describe processes by various levels of granularity, the meaning of model elements is established on a case-by-case basis by machine-readable annotations that link them to entries in public web resources. Annotations may, for instance, relate cell compartments to Gene Ontology entries (Ashburner *et al*, 2000) and small chemical compounds to entries from ChEBI (Degtyarenko *et al*, 2008). The MIRIAM initiative (Le Novère *et al*, 2005) has proposed a standard format for biochemical annotations, consisting of the URN of a web resource, an ID for the referenced resource element, and a biological qualifier stating a logical relation (e.g., 'hasPart') between the model element and the resource entry. The use of public web resources and ontologies like the Systems Biology Ontology makes annotations unambiguous and facilitates search, visualisation, and automated reasoning.

While building complex models, Systems Biologists need to search for relevant models, to rank or classify them, and to check how models differ, overlap, and complement each other (Liebermeister, 2008; Krause *et al*, 2010). Furthermore, models need to be validated and refined with experimental data, which have to be retrieved and aligned to the models beforehand. All these tasks call for automation and most of them require quantitative similarity measures between models and data sets, which should capture their biological meaning and be computable fast and reliably. The situation is comparable to the early days of bioinformatics, when nucleotide sequences became harder and harder to compare until filtering methods like FASTA and BLAST (Lipman and Pearson, 1985; Altschul *et al*, 1990) led to a breakthrough.

A comparison of models can be based on various biological or mathematical aspects, including the biological entities and processes described, the mathematical formalism, details of the equations and numerical values, or even the dynamic behaviour. In particular, most Systems Biology models

describe biological networks and can be depicted and treated as networks themselves. The automatic comparison of biological networks has been widely discussed in the literature. In their review, Sharan and Ideker (2006) distinguish between three main groups of applications. (i) *Network alignment* describes the process in which complete networks (e.g., protein–protein interaction networks from different species; Matthews *et al*, 2001) are compared in order to unveil similar and different regions. (ii) *Network integration* combines networks of different types to gain particular information (e.g., enriched protein interconnection patterns; Zhang *et al*, 2005). (iii) *Network querying* detects parts of a large network that resemble a query motif (e.g., to search how a metabolic pathway is conserved across species; Pinter *et al*, 2005). On the computational side, the alignment and the querying problem can be tackled by similar algorithms, which differ in how they compare network structures and how they relate nodes. The alignment of structures has evolved from simple paths (similar to sequence alignment) (Kelley *et al*, 2003) and trees (Pinter *et al*, 2005) to general graph structures (Yang and Sze, 2007). Depending on the application, the comparison of nodes can be based on their labels, on the relatedness of their annotations (e.g., their EC numbers; Tohsato *et al*, 2000), or on their chemical structures (Hattori *et al*, 2003). For a review on network querying algorithms, the reader is referred to Fionda and Palopoli (2011).

When comparing Systems Biology models, the network structure may be less informative because the same system may be described by alternative models at different levels of granularity (Markevich *et al*, 2004). Graph reduction techniques (Gay *et al*, 2010) can partially handle this problem, but only if the networks are not too different. A more direct way to find biologically similar models, especially for searching and ranking (Henkel *et al*, 2010), is to compare their semantic annotations using methods from information retrieval (IR), as introduced in Box 1. However, the comparison of semantic annotations involves two general challenges: (i) annotations may describe the same chemical entity or process, but point to entries in different web resources; (ii) different web resource entries can share subtle biochemical relationships (e.g., the molecular species ATP<sup>3+</sup> referenced in one model being a special case of—rather than identical to—ATP referenced in another data set). To overcome this problem, intra-ontology relationships and cross references for a large number of relevant web resources have to be combined in an integrative ontology, which can then be used to compare entries from various resources.

## Case study: semantic similarity measures for SBML models

As an example case for the use of semantic annotations, we present a system for retrieval, clustering, and alignment of SBML models. It relies on a technical infrastructure for handling biological concepts (BCs) and on semantics-based similarity scores for models and data sets. We applied our framework to models from BioModels Database, validated the calculated model similarities with human expert knowledge,

### Box 1 Overview of key concepts for the calculation of similarity measures between semantically enriched data

**Information retrieval** Information retrieval (IR) is a field of research investigating how to find relevant documents, or parts of them, in a document resource. Unlike queries in data retrieval, IR queries do not need to follow a fixed format, but can be formulated in natural language and may be fuzzy and incomplete. Since the aim is not to find exact matches, but documents that are relevant for the user, the approaches of IR are often probabilistic and are based on heuristic similarity scores, which summarise the resemblance between the user's request and a document in question. Methods used in IR can differ in four central points. How are query and documents represented (e.g., as sets of terms or as vectors; Salton, 1971)? How are terms interrelated? How are different terms weighted (e.g., by the term frequency–inverse document frequency approach (Jones, 1972) which compares the frequency of a term in one document with the number of documents in which it appears)? And how is this information combined using similarity measures (e.g., the cosine measure; Salton and McGill, 1986)?

**Semantic web** The term 'semantic web' expresses the idea to add explicit, computationally accessible semantic information to documents in order to facilitate automated reasoning. The Resource Description Framework employs triplets of the form (subject, predicate, and object). A statement 'glucose is a sugar', for instance, could be expressed by a triplet ('glucose', 'is\_a', 'sugar'), where each of the three items is taken from a controlled vocabulary or an ontology. Standards for minimal information in machine-readable formats, e.g., MIAME (Brazma *et al*, 2001) enable computers to assess the content of a data set or model.

**Similarities between ontology terms** An ontology defines a set of concepts and interconnects these concepts by various types of relationships. Early similarity measures between ontology elements were defined based on the hierarchy of 'is\_a' relations and counted the number of edges on the shortest path between two concepts (Rada *et al*, 1989). Other measures defined the similarity by the highest information content among the super-concepts subsuming both compared concepts (e.g., 'animal' subsuming the concepts 'cat' and 'fish'). To compute information the measures make use of the knowledge from a so-called corpus, e.g., a long text, in which the appearance of different concepts can be counted (Resnik, 1995; Lin, 1998). If specialisations of a concept are rarely used, the similarity between its subconcepts is increased. This idea has also been used to reweight edges in an ontology, which can improve the performance of edge-based methods (Jiang and Conrath, 1997). The quality of different similarity measures and their combinations has been rigorously assessed in Li *et al* (2003).

and present a number of practical applications. Researchers can use our online services at <http://www.semanticsbml.org> to retrieve Systems Biology models resembling a given SBML model or related to an experimental data set. Furthermore, they can cluster models by their semantic similarities and visually align their elements. The mathematical and technical details are explained in parts below and extensively in the Supplementary Appendix.

## Challenges in the automatic comparison of model elements

The biological meaning of SBML elements can be declared by annotations according to the MIRIAM standard. Comparing SBML elements comprising possibly many annotations imposes even more challenges than the comparison of single annotations: (i) the relationships between model elements and resource entries, stated by qualifiers, may be complex (e.g., 'hasPart' rather than a simple 'is'); (ii) each model element may contain several annotations, describing its different aspects; (iii) annotations may be missing, unspecific, or simply wrong.

### Semantic annotations in BioModels database

To compare biological identifiers and to evaluate the relationships between them efficiently, we developed the query engine libSBAnnotation, which collects biochemical knowledge from several public web resources and combines it in a single ontology. Equivalent entries from different resources are replaced by single ‘Biological Concepts’ (BCs), whereas similar entries (e.g., ‘ $\alpha$ -D-glucose’ versus ‘glucose’) are represented by different BCs, but connected by ontology relationships (e.g., ‘is\_a’). Queries can be posed through a programming interface or through a web service compliant to the Representational State Transfer (REST) software architecture style (Fielding, 2000).

The libSBAnnotation makes it easy to explore the semantic annotations present in BioModels Database, the major public collection of MIRIAM-compliant models (249 models in the 17th release from May 2010, which we use in the current study). A comprehensive statistics for the most recent

BioModels release is provided at our online service. Approximately 69% of all compartments, species, and reactions are annotated. They show about 1.7 annotations per annotated element and almost all of them carry ‘isVersionOf’, ‘hasPart’, or ‘is’ qualifiers. The annotations refer to a broad range of web resources and the high abundance of Gene Ontology and UniProt entries shows that the models contain more proteins than small metabolites.

Figure 1 shows the prevalence of about 2000 BCs within all 249 models in the form of an annotation matrix. Positive matrix elements indicate which models (columns) contain annotations referring to certain BCs (rows). The numerical values may also state how often a BC appears in a model and which qualifiers are used. The matrix columns, called feature vectors, can be seen as the ‘annotation fingerprints’ of models and may serve for simple comparisons and visualisation by multivariate statistical methods.



**Figure 1** Annotation matrix. Semantic annotations link the elements of Systems Biology models to Biological Concepts from public web resources. The associations between them can be represented by a matrix: positive entries (red) show that a model (column) contains an annotation pointing to a certain Biological Concept (row). Left: annotation matrix for the BioModels Database, sorted by two-way agglomerative clustering. Right: close-up showing a number of MAP kinase models and Biological Concepts. Matrix visualised by GenePattern (Reich *et al*, 2006).



## Similarity measures for BCs, annotations, and models

The libSBAnnotation interconnects BCs by various semantic relationships and thus forms an ontology. To express the direct and indirect relationships between BCs by numbers, we developed a series of similarity measures which resemble the scores used in semantic text analysis. Based on the similarities between individual BCs, we then define similarity measures between entire models. We investigated two groups of such measures and tested their performance in practical applications.

1. Vector-based similarity measures are solely based on the feature vectors, i.e., on the set of BCs referenced by a model. Two feature vectors are compared by the cosine coefficient and a special metric is used to acknowledge that annotations can point to different, yet similar BCs (e.g., ‘ $\alpha$ -D-glucose’ versus ‘glucose’).
2. Structure-based similarity measures, in contrast, start with a pairwise comparison of individual model elements and combine the resulting similarities in more complex model similarity scores. In the spirit of probabilistic reasoning, one of these similarity measures can combine evidence from several annotations, distinguish between missing information (i.e., no element annotations) and negative information (i.e., annotations pointing to different BCs), and weight different combinations of biological qualifiers and relationships between BCs.

For a practical test, we evaluated all similarity measures with benchmark models from BioModels Database, the largest public collection of curated SBML models. After manually classifying the models by the biochemical pathways described, we clustered them all by each of the measures (similar to Figure 1) and compared the clusters with the predefined classification. A detailed evaluation can be found in the Supplementary Appendix.

The vector-based similarity measures have a number of advantages: first of all, they performed well in the comparison and are easy to compute. Moreover, being solely based on sets of annotations, these annotation measures not only apply to kinetic or structural SBML models, but also to annotated ‘omics’ data or any type of data associated with a list of BCs. Therefore, they are used in our online tools and will be explained below. More details and descriptions of other similarity measures are given in the Supplementary Appendix.

### Similarity between BCs

Similarity measures for ontology elements have been discussed in the literature from a theoretical (Lin, 1998) and a practical point of view (Resnik, 1995; Budanitsky and Hirst, 2001; Li *et al*, 2003) and have been implemented in software tools (Lord *et al*, 2003). They are usually computed from the distance between entries in the relationship graph, their most specific common ancestor, and a corpus, a collection of text or data in which the appearance frequencies of ontology terms can be counted. Following Li *et al*, we define a similarity  $\sigma$  between BCs  $\mu$  and  $\nu$  taking into account three factors: (i) their weighted distance ( $f_1$ ) in the ontology forest, (ii) their

depths ( $f_2$ ) (distance from a root), and (iii) their rarity ( $f_3$ ) in BioModels Database. The three factors are combined by the formula

$$\sigma_{BC}(\mu, \nu) = f_1(\mu, \nu)^{f_2(\mu, \nu)} \times f_3(\mu) \times f_3(\nu).$$

The factor  $f_1$  yields a high similarity if two BCs are connected by a short relationship path

$$f_1(\mu, \nu) = \max_{p \in P} \prod_{r \in p} f_{rts}(r),$$

where  $p \in P$  is a possible path of relation arrows ( $r$ ) between the two BCs and  $f_{rts}$  scores each relation type by a value between 0 and 1 (see Supplementary Appendix for numerical values). If there is no path between the BCs,  $f_1$  is set to 0. A sensitivity analysis in the Supplementary Appendix shows that the choice of numerical values for  $f_{rts}$  has only little effect on the model retrieval results. Since too few benchmark examples are available to optimise these parameters reliably, we use *ad hoc* values chosen before testing any of the measures.

BCs that are deeper in the relationship graph are usually more precise because they comprise fewer subconcepts (e.g., ‘D-glucopyranose’ being a subconcept of ‘carbohydrate’). Accordingly, if two pairs of BCs are connected by a similar path, the pair with the lower depth (e.g., ‘carbohydrate’ and ‘sugar’) should be less similar than the more specific pair (e.g., ‘D-glucopyranose’ and ‘ $\alpha$ -D-glucose’). This is implemented by the exponent

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

where  $d(\mu)$  is the path length of ‘is\_a’ relationships between an ontology element  $\mu$  and its root.

Since some BCs (e.g., ATP) appear very often in BioModels Database, a semantic density factor  $f_3$  can be introduced to downweight them in the similarity measure. For each BC  $\mu$ , it reads

$$f_3(\mu) = 1 - \frac{\log(c_\mu^c + 1)}{\log c_\Omega},$$

where  $c_\mu$  is the number of occurrences of  $\mu$  in BioModels Database,  $c_\mu^c = c_\mu + \sum_{\xi \in \text{children}(\mu)} c_\xi^c$  is the number of  $\mu$  and all its ‘is\_a’ specialisations, and  $c_\Omega = 1 + \sum_{\forall \xi: c_\xi > 0} (c_\xi + 1)$  is a normalisation term. However, the term  $f_3$  did not seem to improve the results in our evaluation. This agrees with the findings of Li *et al* (2003), who concluded that the use of an ‘information factor’ in text analysis decreases the quality of their similarity measure. Because of this and because the frequency of individual annotations is already part of our null model for computing  $P$ -values, we omitted this term from the online model search.

### Vector-based model similarity

The elements of an SBML model, such as cellular compartments, molecular species, or biochemical reactions, can be annotated with links to various web resources. Given a general list of BCs, each model or data set  $M$  can be represented by a feature vector  $v_M$  with components  $v_{iM}=1$  if the  $i$ th BC appears in an annotation in model  $M$  and  $v_{iM}=0$  otherwise. Similarities between two models  $M$  and  $N$  can be defined by functions of their feature vectors. From the various measures used in IR, we

chose the cosine of the angle between the feature vectors (van Rijsbergen, 1979; Salton and McGill, 1986). Despite its simplicity, this cosine coefficient allows for reasonable comparisons between models. However, it cannot detect the resemblance between similar, but slightly different BCs (e.g., CHEBI:17634 for D-glucose and CHEBI:17925 for  $\alpha$ -D-glucose). To capture such biochemical similarities, we replace the scalar product by a quadratic form based on the similarity matrix  $S$  for the BCs ( $S_{ij} = \sigma_{BC}(\mu, \nu)$ , where  $\mu$  is the  $i$ th and  $\nu$  is the  $j$ th BC):

$$\sigma_{Mo}(M, N) = \frac{v_M^T S v_N}{\sqrt{v_M^T S v_M} \sqrt{v_N^T S v_N}}$$

Since the feature vectors and the similarities  $\sigma_{BC}(\mu, \nu)$  are non-negative, this formula yields non-negative model similarities even if the similarity matrix is not positive definite. For positive-definite matrices  $S$ , the formula can be interpreted in terms of transformed feature vectors as proposed in the topic-based vector space model (TVSM; Becker and Kuroopka, 2003).

### Implementation and data

The query engine libSBAnnotation and all described methods were implemented in Python. The code is freely available at <http://sourceforge.net/projects/semanticsbml>. Online tools for similarity calculations, model retrieval, clustering based on the TVSM similarities, and alignment based on greedy pairing, as well as a public REST API for programmatic access are provided at <http://www.semanticsbml.org>.

### Model search

#### Ranked retrieval of SBML models from BioModels database






















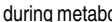
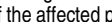
Model similarities can be used to find models or data sets referring to a given query model or a list of BCs. As a practical

application, we have implemented a similarity search for models from BioModels Database (Figures 2 and 3). The retrieved models are ranked by similarity scores, where high scores indicate that query model and retrieved model share a large fraction of similar annotations. To discard models describing unrelated pathways, we filter out low-scoring search results by a statistical significance test. Our null hypothesis states that BCs appear in the models independently and with the same BC-specific frequencies as in BioModels Database. Accordingly, low  $P$ -values indicate that the query model and a retrieved model share a set of common BCs that is unlikely to appear just by chance, which suggests that they describe the same biological pathways.

Given the similarity score  $\sigma$  of a certain retrieved model, the  $P$ -value states how probable it is to obtain an equal or higher similarity score from a random model. For the calculation, we randomly sample feature vectors in which each BC appears with the probability  $(b + 1)/(B + 1)$ , where  $b$  is the number of models referring to this BC and  $B$  is the total number of models. We generate  $N=998$  such random models, check how many of them show higher similarities to the query model than our actual retrieved model. From this number  $n$ , the  $P$ -value is estimated by the Bayesian estimator  $\langle P \rangle = (n + 1)/(N + 2)$  with a uniform prior for the  $P$ -value. For practical reasons, we also compute a second  $P$ -value for the model overlap  $v_M^T v_N$ , which can be computed without the need for random sampling. Analytically calculating  $P$ -values for other similarity scores turned out to be too slow for efficient online searches. More details on the methods can be found in the Supplementary Appendix.

#### Search for models related to an experimental result

A model search may begin with a list of genes involved in a certain biological process. As an example (see Figure 2), we considered a microarray study on gene regulation during

Model	BioModel	Similarity	P-value	Overlap	P-value	
Wolf2001_respiratory_oscillations	BIOMD0000000090	0.207	<=1e-3	6	6.6e-09	
Chassagnole2001_Threonine_Synthesis	BIOMD0000000066	0.184	<=1e-3	4	1.5e-05	
Curien2009_Aspartate_Metabolism	BIOMD0000000212	0.170	<=1e-3	5	3.6e-07	
Curien2003_MetThr_synthesis	BIOMD0000000068	0.141	<=1e-3	2	1.0e-02	
Proctor2007_ubiquitine	BIOMD0000000105	0.098	2.0e-03	1	1.4e-01	
Curto1998_purineMetabol	BIOMD0000000015	0.063	1.1e-02	2	1.0e-02	
Ibrahim2008_Spindle_Assembly_Checkpoint_dissociation	BIOMD0000000186	0.057	1.8e-02	0	1.0e+00	
Ibrahim2008_Spindle_Assembly_Checkpoint_convey	BIOMD0000000187	0.057	1.8e-02	0	1.0e+00	
Rodriguez-Caso2006_Polyamine_Metabolism	BIOMD0000000190	0.040	7.1e-02	1	1.4e-01	
Nijhout2004_Folate_Cycle	BIOMD0000000213	0.032	1.1e-01	1	1.4e-01	
Morrison1989_FolateCycle	BIOMD0000000018	0.030	1.3e-01	1	1.4e-01	
Zatorsky2006_p53_Model3	BIOMD0000000154	0.023	2.5e-01	0	1.0e+00	
Zatorsky2006_p53_Model6	BIOMD0000000155	0.023	2.5e-01	0	1.0e+00	
Hunziker2010_p53_StressSpecificResponse	BIOMD0000000252	0.023	2.5e-01	0	1.0e+00	
Zatorsky2006_p53_Model5	BIOMD0000000156	0.022	2.7e-01	0	1.0e+00	
Zatorsky2006_p53_Model4	BIOMD0000000157	0.022	2.7e-01	0	1.0e+00	
Zatorsky2006_p53_Model2	BIOMD0000000158	0.022	2.7e-01	0	1.0e+00	
Zatorsky2006_p53_Model1	BIOMD0000000159	0.022	2.7e-01	0	1.0e+00	
Proctor2008_p53_Mdm2_ATM	BIOMD0000000188	0.013	4.3e-01	0	1.0e+00	
McClellan2007_CrossTalk	BIOMD0000000116	0.012	4.7e-01	0	1.0e+00	
Proctor2008_p53_Mdm2_ARF	BIOMD0000000189	0.012	4.9e-01	0	1.0e+00	
Haberichter2007_cellcycle	BIOMD0000000109	0.011	5.0e-01	0	1.0e+00	
Sasagawa2005_MAPK	BIOMD0000000049	0.006	5.5e-01	0	1.0e+00	

**Figure 2** Model search based on a microarray study. The experiment (Klevecz et al, 2004) revealed differential gene expression during metabolic oscillations in yeast, which are coupled with bursts in DNA replication. Using the list of differentially expressed genes as a query, we obtained models of the affected pathways. The retrieved models describe methionine or more general amino-acid metabolism (BioModels 66, 212, 68, 15, 190, 213, and 18), sulphur metabolism (90), ubiquitination (105), and the DNA polymerase (15), and thus cover three of the six functional categories of the query genes. Bar lengths and colours show the vector-based model similarity scores.

Model	BioModel	Similarity	P-value	Overlap	P-value
Huang1996_MAPK_ultrasens	BIOMD0000000009	1.000	<=1e-3	30	0.0e+00
Levchenko2000_MAPK_noScaffold	BIOMD0000000011	0.930	<=1e-3	28	0.0e+00
Levchenko2000_MAPK_Scaffold	BIOMD0000000014	0.874	<=1e-3	26	0.0e+00
Kholodenko2000_MAPK_feedback	BIOMD0000000010	0.830	<=1e-3	20	0.0e+00
Markevich2004_MAPK_orderedElementary	BIOMD0000000026	0.749	<=1e-3	16	2.9e-15
Markevich2004_MAPK_phosphoRandomElementary	BIOMD0000000028	0.692	<=1e-3	15	9.1e-14
Markevich2004_MAPK_AllRandomElementary	BIOMD0000000030	0.692	<=1e-3	15	9.1e-14
Markevich2004_MAPK_orderedMM	BIOMD0000000027	0.691	<=1e-3	12	9.8e-10
Markevich2004_MAPK_orderedMM2kinases	BIOMD0000000031	0.691	<=1e-3	12	9.8e-10
Markevich2004_MAPK_phosphoRandomMM	BIOMD0000000029	0.626	<=1e-3	11	1.6e-08
Hornberg2005_ERKcascade	BIOMD0000000084	0.523	<=1e-3	9	2.7e-06
McClean2007_CrossTalk	BIOMD0000000116	0.453	<=1e-3	8	2.7e-05
Kofahl2004_pheromone	BIOMD0000000032	0.441	<=1e-3	12	9.8e-10
Goldbeter1991_MinMitOscil_ExpInact	BIOMD0000000004	0.389	<=1e-3	3	1.5e-01
Brown2004_NGF_EGF_signaling	BIOMD0000000033	0.371	<=1e-3	9	2.7e-06
Ung2008_EGFR_Endocytosis	BIOMD0000000205	0.363	<=1e-3	8	2.7e-05
Kim2007_Wnt_ERK_Crosstalk	BIOMD0000000149	0.355	<=1e-3	10	2.2e-07
Goldbeter1991_MinMitOscil	BIOMD0000000003	0.349	<=1e-3	3	1.5e-01
Sasagawa2005_MAPK	BIOMD0000000049	0.339	<=1e-3	9	2.7e-06
Swat2004_Mammalian_G1_S_Transition	BIOMD0000000228	0.317	<=1e-3	2	4.0e-01
Tyson1991_CellCycle_6var	BIOMD0000000005	0.304	<=1e-3	4	4.2e-02
Goldbeter1995_CircClock	BIOMD0000000016	0.274	<=1e-3	4	4.2e-02
Novak1997_CellCycle	BIOMD0000000007	0.259	<=1e-3	1	7.6e-01
Novak2001_FissionYeast_CellCycle	BIOMD0000000111	0.255	<=1e-3	2	4.0e-01
Leloup1999_CircClock	BIOMD0000000021	0.246	<=1e-3	4	4.2e-02
Birtwistle2007_ErbB_Signalling	BIOMD0000000175	0.236	<=1e-3	1	7.6e-01
Neves2008_Cell_Shape	BIOMD0000000182	0.222	<=1e-3	6	1.6e-03
Leloup1998_CircClock_LD	BIOMD0000000171	0.219	<=1e-3	4	4.2e-02
Veening2008_DegU_Regulation	BIOMD0000000240	0.211	4.0e-03	2	4.0e-01
Chen2004_CellCycle	BIOMD0000000056	0.209	4.0e-03	4	4.2e-02
Fernandez2006_ModelA	BIOMD0000000152	0.207	5.0e-03	2	4.0e-01
Fisher2006_NFAT_Activation	BIOMD0000000123	0.199	7.0e-03	2	4.0e-01
Hatakeyama2003_MAPK	BIOMD0000000146	0.198	7.0e-03	1	7.6e-01

**Figure 3** Results of a semantic model search in BioModels Database. Starting from the kinase cascade model of Huang and Ferrell (1996), a ranked list of similar models was retrieved automatically. The first 15 models contain complete kinase cascades or parts of them. The top hit is the query model itself.

metabolic oscillations in the yeast *S. cerevisiae*. The experimental data of Klevecz *et al* (2004) show that the observed oscillations are coupled with bursts in DNA replication. Differentially expressed genes tend to be involved in sulphur and methionine metabolism and in the production of ubiquitine proteasomes, ribosomes, and the DNA polymerase. To retrieve models related to this gene set, we described the genes by MIRIAM-compliant annotations and started a search for relevant models. The retrieval returned models of methionine and sulphur metabolism, a model of the ubiquitine proteasome system, and a model containing a DNA polymerisation reaction (see Figure 2). Although the functional categories of the query genes were not explicitly used in the query, they perfectly agree with the search results. Similar model searches could start from any list of metabolites, genes, or proteins. Further examples and practical hints for the annotation and retrieval process can be found on our website.

### Search results for a signal transduction model

The result of a model search starting from an MAP kinase cascade model, BioModel 9 (Huang and Ferrell, 1996), is shown in Figure 3. The topmost 15 models in the list describe either MAP kinase cascades or parts of them. While the models 11, 14, and 10 are as detailed as the query model, the models of Markevich (26–31) represent only the activation of MAPK, but in more detail, and the models 84, 116, 32, 149, and 33 contain additional proteins around the MAP kinase cascade. All similarities are highly significant (estimated *P*-values around  $10^{-3}$ ). Models further down in the list share some general annotations with the query model, for instance, Gene Ontology terms for protein phosphorylation and dephosphorylation, but they rarely describe MAP kinase cascades. Depending on the frequency of the common annotations, the

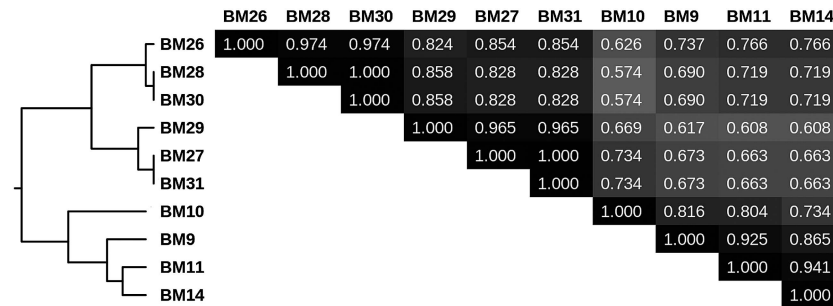
retrieved models may still appear significant, but they show much lower similarity scores than the first 15 hits.

### Model clustering

Unsupervised clustering is one of the prominent applications of similarity measures. As an example, we clustered the first 10 models from Figure 3 by agglomerative clustering using vector-based similarities. As shown in Figure 4, the two model groups describing either the complete MAP kinase cascade (9, 10, 11, and 14) or MAPK activation (26–31) are clearly distinguished. Furthermore, the models 11 and 14, which stem from the same publication (Levchenko *et al*, 2000), show the highest similarity among the complete MAP kinase cascades, whereas model 10, the only model with enzymatic reactions represented with Michaelis–Menten-like kinetics, appears most distant to all others. Among the MAPK activation models, the clustering clearly distinguishes between models containing effective enzymatic rate laws (27, 29, and 31) and the ones containing elementary reaction steps (26, 28, and 30). The reason for this distinction is not the structural difference between the models, but the fact that many elementary reactions in these models (in contrast to the enzymatic ones) were annotated with Gene Ontology terms for enzyme binding or dissociation.

### Model alignment

One of the key challenges in automated model merging is to match equivalent elements from two models. To realise such a model alignment, we employed a greedy pairing: the two elements with the highest pairwise similarity are successively matched until all remaining similarities fall below a certain threshold. At our website, the user can visually align



**Figure 4** Clustering of computational models. The MAP kinase models retrieved from BioModels Database (see Figure 3) contain subgroups that are successfully detected by the clustering. Despite their different network structures, the models of Markevich (26–31), describing the same biological pathway, show high pairwise similarities ( $> 0.82$ ). Dendrogram drawn by DendroUPGMA (Garcia-Valve *et al*, 1999).

annotated SBML models or data sets to similar models from BioModels Database. An example, the visual alignment between the MAP kinase cascade models, BioModel 84 and BioModel 9, is shown in Figure 5. Although both models share the same general structure, the Huang model shows the phosphorylation states of the MAP kinases in much higher resolution. Structure-based alignment methods could not detect the similarity between these models without heavily increasing the ‘fuzziness’ (number of node insertions/deletions/mismatches) of their matching, which in turn leads to a low specificity.

As a further application, we used model alignments to tile the metabolic network of the yeast *S. cerevisiae* with kinetic models available in BioModels Database. By iteratively aligning kinetic models to the yeast consensus metabolic network (Herrgård *et al*, 2008), we could cover about 15% of the network with eight kinetic models (most of which described the metabolism of other organisms), whereas all other models contributed only few additional elements (see the Supplementary Appendix for details). This automatic comparison shows that existing kinetic models would by far not suffice to build a comprehensive model of yeast metabolism. In the future, the same method could help to find white spots in the cellular networks that might deserve further modelling efforts.

## Discussion

The construction of large Systems Biology models in a bottom-up style requires models that are easy to reuse. Public model repositories, standard formats, and annotation schemes have already been established, and the wealth of information stored in models is ready to be processed by computer programmes. While interconversion of annotations is mainly a matter of technology, defining suitable similarities between models is a more delicate task. The reason is that similarity measures are not given *a priori*, but need to reflect specific human intentions and expectations in order to be useful for practical applications.

Computational models can resemble each other in two complementary ways: first, they may describe similar biological systems; and second, in case they do, they may describe them using a similar level of granularity, similar formulas, or similar quantitative values. In the present approach, we

focused on the first aspect, which is fully captured by the biological annotations. The technical challenges mentioned above were solved by interconnecting model annotations and BCs, by assigning quantitative weights to the biological qualifiers and relationships between BCs, and by condensing all information within the similarity measures.

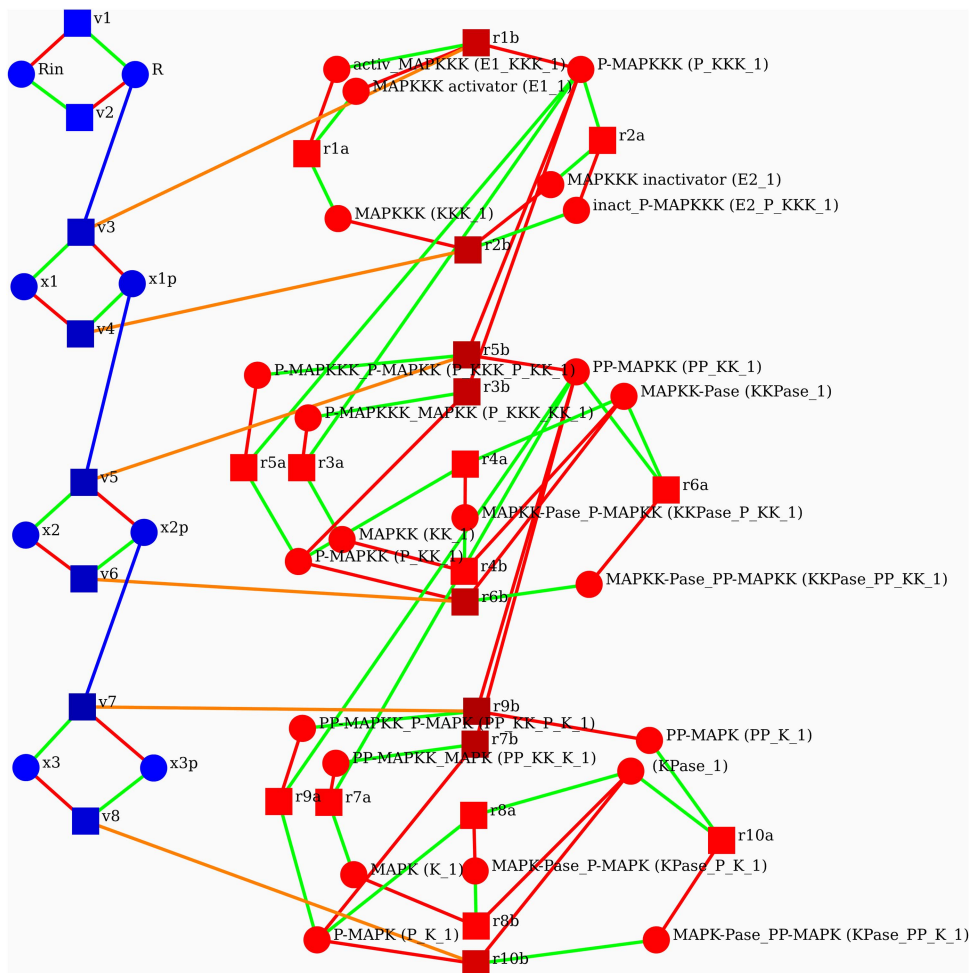
The second aspect, which concerns model formulation, network structure, mathematical statements, and numerical values, was ignored here. Of course, the similarity measures could be extended to compare enzymatic rate laws (e.g., by evaluating annotations with Systems Biology Ontology identifiers) or mathematical formulae. However, similarity scores for mathematical statements or network structures would strongly depend on specific model formalisms, while the comparison of annotations is not even limited to Systems Biology models, but may apply to models from other fields and even experimental data sets or annotated scientific literature (Cheung *et al*, 2010).

Another reason for semantic comparisons is that biologists typically search for models describing a certain biochemical process, irrespective of the mathematical details. A paper such as Markevich *et al* (2004), for instance, represents the same biochemical process (MAPK phosphorylation) by six different mathematical structures and, therefore, different reaction networks. Annotations make it easy to recognise the similarity between these models, while network-based model similarities would emphasise their differences.

Like many other classification tasks, model retrieval crucially depends on a sensible choice of the null model used for computing the *P*-values. Since the null model is used to distinguish between meaningless and meaningful similarities, it needs to be chosen as carefully as the similarity measure itself. In general, it should capture the typical properties of models that are not specifically interesting as search results. In the present approach, the main aim was to find models that specifically share annotations with a query model. Unspecific BCs, especially those that are very frequent, are likely to lead to spurious similarity values. Our null model was tailored to reproduce exactly this effect in order to tag the resulting low similarities as insignificant. In the future, larger model databases and more specific search tasks may raise the need for advanced null models, specifically constructed to match and exclude unintended search results.

The comparison of SBML models by semantic annotations works well in practise and may pave the way to promising

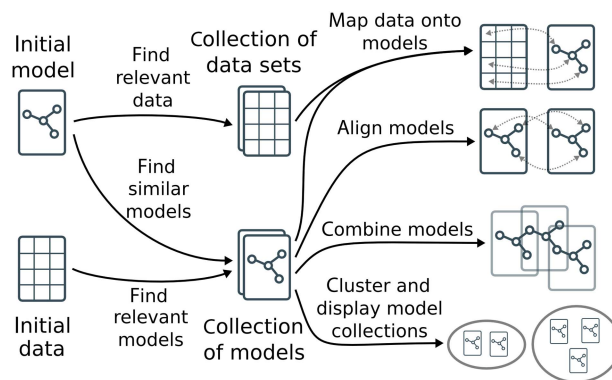




**Figure 5** Visual alignment between computational models. An MAP kinase model (BioModel 84; Hornberg *et al*, 2005) (blue) is aligned with the more detailed BioModel 9 (Huang and Ferrell, 1996) (red). The reaction networks represent chemical species (circles) and reactions (squares) connected by reactant (green) and product edges (red). Orange edges connect elements between models if their similarity scores exceed a threshold value of 0.25.

applications. By transforming and normalising the semantic feature vectors, similarities can be rewritten in terms of Euclidean distances, which makes them amenable to multivariate methods such as Kohonen maps, biclustering, principal and independent component analysis (Pearson, 1901; Comon, 1994), non-negative matrix factorisations (Lee and Seung, 1999), classification by support vector machines, and search for prototype models. These methods, in turn, may have various practical applications in the visualisation and statistical analysis of large model sets.

As depicted in Figure 6, automated searches for models and experimental data can be helpful in early and later stages of the modelling process. Existing models can provide information about additional reactions, enzymatic rate laws, and parameter values, or suggest alternative descriptions of biochemical processes. More complex searches using positive and negative weights for the individual features, e.g., for models that contain certain annotations and lack certain others, could help to extend existing models by additional pathways. Finally, the possibility to start the retrieval process from ‘omics’ data opens up new applications, including pathway enrichment



**Figure 6** Semantic model comparison can be useful during hypotheses generation, modelling, experimental verification, and model refinement. Given a model or an experimental data set, similar models or data can be found in repositories and be used to extend existing models, refine them using data, and finally select the most appropriate model. Models and data sets of interest can further be mapped, aligned, combined, and classified or displayed by clustering.

analyses, comparison between experimental data and simulation results, or automated model parameter fitting and model selection.



## Conclusion

As models and data in Systems Biology are rapidly accumulating, automatic searches for models or data sets and pairwise alignments between them become increasingly important. For efficient searches, models and data have to adhere to standard formats, contain reliable biological annotations, and be stored in central, publicly accessible repositories. Public databases already provide a significant number of well-annotated models and data, and model comparison may promote various applications, allowing to exploit an otherwise hardly manageable amount of knowledge. Facilitating the reuse of models and data, such comparisons may become a basic method in computational Systems Biology, just as tools like BLAST (Altschul et al, 1990) became to scientists dealing with sequence data.

## Supplementary information

Supplementary information is available at the *Molecular Systems Biology* website ([www.nature.com/msb](http://www.nature.com/msb)).

## Acknowledgements

We are grateful to Dagmar Waltemath and Ron Henkel for lively discussions and helpful comments on the manuscript. This work was supported by the British Biotechnology and Biological Sciences Research Council [BB/F010516/1] (to NLN), the International Max Planck Research School for Computational Biology and Scientific Computing, the German Research Foundation [CRC 618], the BMBF SysMO Project Translucent2 [contract number 0315786A], and the European Commission [BaSysBio, grant number LSHG-CT-2006-037469] (to EK).

*Author contributions:* WL, MS, and FK designed research; MS and FK implemented methods; MS analysed data; and MS, WL, FK, NLN, and EK wrote the paper.

## Conflict of interest

The authors declare that they have no conflict of interest.

## References

- Altschul S, Gish W, Miller W, Myers E, Lipman D (1990) Basic local alignment search tool. *J Mol Biol* **215**: 403–410
- Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry J, Davis A, Dolinski K, Dwight S, Eppig J, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. *Nat Genet* **25**: 25–29
- Bader G, Cary M, Sander C (2006) Pathguide: a pathway resource list. *Nucleic Acids Res* **34**: D504
- Becker J, Kuropka D (2003) Topic-based vector space model. In: *Proceedings of the Sixth International Conference on Business Information Systems*, pp 7–12
- Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball C, Causton H, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S et al (2001) Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nat Genet* **29**: 365–371
- Budanitsky A, Hirst G (2001) Semantic distance in WordNet: an experimental, application-oriented evaluation of five measures. In: *Workshop on WordNet and Other Lexical Resources*, Vol. 2
- Cheung K, Samwald M, Auerbach R, Gerstein M (2010) Structured digital tables on the Semantic Web: toward a structured digital literature. *Mol Syst Biol* **6**: 403
- Comon P (1994) Independent component analysis, a new concept? *Signal Process* **36**: 287–314
- Degtyarenko K, de Matos P, Ennis M, Hastings J, Zbinden M, McNaught A, Alcantara R, Darsow M, Guedj M, Ashburner M (2008) ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res* **36**: D344
- Fielding R (2000) *Architectural styles and the design of network-based software architectures*. Ph.D. thesis, Irvine: University of California
- Fionda V, Palopoli L (2011) Biological network querying techniques: analysis and comparison. *J Comput Biol* **18**: 595–625
- Garcia-Vallve S, Palau J, Romeu A (1999) Horizontal gene transfer in glycosyl hydrolases inferred from codon usage in *Escherichia coli* and *Bacillus subtilis*. *Mol Biol Evol* **16**: 1125
- Gay S, Soliman S, Fages F (2010) A graphical method for reducing and relating models in systems biology. *Bioinformatics* **26**: i575
- Hattori M, Okuno Y, Goto S, Kanehisa M (2003) Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J Am Chem Soc* **125**: 11853–11865
- Henkel R, Endler L, Le Novère N, Peters A, Waltemath D (2010) Ranked retrieval of computational biology models. *BMC Bioinformatics* **11**: 423
- Herrgård M, Swainston N, Dobson P, Dunn W, Arvas K, Arvas M, Bütthgen N, Borger S, Costenoble R, Heinemann M, Hucka M, Le Novère N, Li P, Liebermeister W, Mo M, Oliveira A, Petranovic D, Pettifer S, Simeonidis E, Smallbone K et al (2008) A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nat Biotechnol* **26**: 1155–1160
- Hornberg J, Bruggeman F, Binder B, Geest C, de Vaate A, Lankelma J, Heinrich R, Westerhoff H (2005) Principles behind the multifarious control of signal transduction. *FEBS J* **272**: 244–258
- Huang C, Ferrell J (1996) Ultrasensitivity in the mitogen-activated protein kinase cascade. *Proc Natl Acad Sci USA* **93**: 10078
- Hucka M, Finney A, Sauro H, Bolouri H, Doyle J, Kitano H, Arkin A, Bornstein B, Bray D, Cornish-Bowden A, Cuellar AA, Dronov S, Gilles E, Ginkel M, Gor V, Goryanin II, Hedley WJ, Hodgman TC, Hofmeyr JH, Hunter PJ et al (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19**: 524–531
- Jiang J, Conrath D (1997) Semantic similarity based on corpus statistics and lexical taxonomy. In: *Tenth International Conference on Research on Computational Linguistics (ROCLING X)*
- Jones K (1972) A statistical interpretation of term specificity and its application in retrieval. *J Doc* **28**: 11–21
- Kelley B, Sharan R, Karp R, Sittler T, Root D, Stockwell B, Ideker T (2003) Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc Natl Acad Sci USA* **100**: 11394–11399
- Klevecz R, Bolen J, Forrest G, Murray D (2004) A genomewide oscillation in transcription gates DNA replication and cell cycle. *Proc Natl Acad Sci USA* **101**: 1200
- Krause F, Uhlendorf J, Lubitz T, Schulz M, Klipp E, Liebermeister W (2010) Annotation and merging of SBML models with semantic SBML. *Bioinformatics* **26**: 421
- Le Novère N, Bornstein B, Broicher A, Courtot M, Donizelli M, Dharuri H, Li L, Sauro H, Schilstra M, Shapiro B, Snoep J, Hucka M (2006) BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Res* **34**: D689–D691
- Le Novère N, Finney A, Hucka M, Bhalla U, Campagne F, Collado-Vides J, Crampin E, Halstead M, Klipp E, Mendes P, Nielsen P, Sauro H, Shapiro B, Snoep JL, Spence HD, Wanner BL

- (2005) Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat Biotechnol* **23**: 1509–1515
- Lee D, Seung H (1999) Learning the parts of objects by non-negative matrix factorization. *Nature* **401**: 788–791
- Levchenko A, Bruck J, Sternberg P (2000) Scaffold proteins may biphasically affect the levels of mitogen-activated protein kinase signaling and reduce its threshold properties. *Proc Natl Acad Sci USA* **97**: 5818
- Li Y, Bandar Z, McLean D (2003) An approach for measuring semantic similarity between words using multiple information sources. *IEEE Trans Knowl Data Eng* **15**: 871–882
- Liebermeister W (2008) Validity and combination of biochemical models. In: *Proceedings of Third International ESCEC Workshop on Experimental Standard Conditions on Enzyme Characterizations*
- Lin D (1998) An information-theoretic definition of similarity. In: *Proceedings of the 15th International Conference on Machine Learning*, Vol. 1, pp 296–304
- Lipman D, Pearson W (1985) Rapid and sensitive protein similarity searches. *Science* **227**: 1435
- Lord P, Stevens R, Brass A, Goble C (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics* **19**: 1275
- Markevich N, Hoek J, Kholodenko B (2004) Signaling switches and bistability arising from multisite phosphorylation in protein kinase cascades. *J Cell Biol* **164**: 353
- Matthews L, Vaglio P, Reboul J, Ge H, Davis B, Garrels J, Vincent S, Vidal M (2001) Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or ‘interologs’. *Genome Res* **11**: 2120
- Olivier B, Snoep J (2004) Web-based kinetic modelling using JWS online. *Bioinformatics* **20**: 2143–2144
- Pearson K (1901) On lines and planes of closest fit to systems of points in space. *Philos Mag Series 6* **2**: 559–572
- Pinter R, Rokhlenko O, Yeger-Lotem E, Ziv-Ukelson M (2005) Alignment of metabolic pathways. *Bioinformatics* **21**: 3401
- Rada R, Mili H, Bicknell E, Blettner M (1989) Development and application of a metric on semantic nets. *IEEE Trans Syst Man Cybern* **19**: 17–30
- Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov J (2006) GenePattern 2.0. *Nat Genet* **38**: 500–501
- Resnik P (1995) Using information content to evaluate semantic similarity in a taxonomy. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence*
- Salton G (1971) *The SMART Retrieval System—Experiments in Automatic Document Processing*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.
- Salton G, McGill M (1986) *Introduction to Modern Information Retrieval*. New York: McGraw-Hill, Inc.
- Sharan R, Ideker T (2006) Modeling cellular machinery through biological network comparison. *Nat Biotechnol* **24**: 427–433
- Tohsato Y, Matsuda H, Hashimoto A (2000) A multiple alignment algorithm for metabolic pathway analysis using enzyme hierarchy. In: *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology (ISMB'00)*, pp 376–383
- van Rijsbergen C (1979) *Information Retrieval*, 2nd edn, London: Butterworths
- Yang Q, Sze S (2007) Path matching and graph matching in biological networks. *J Comput Biol* **14**: 56–67
- Zhang L, King O, Wong S, Goldberg D, Tong A, Lesage G, Andrews B, Bussey H, Boone C, Roth F (2005) Motifs, themes and thematic maps of an integrated *Saccharomyces cerevisiae* interaction network. *J Biol* **4**: 6



*Molecular Systems Biology* is an open-access journal published by *European Molecular Biology Organization* and *Nature Publishing Group*. This work is licensed under a Creative Commons Attribution-NonCommercial-Share Alike 3.0 Unported License.