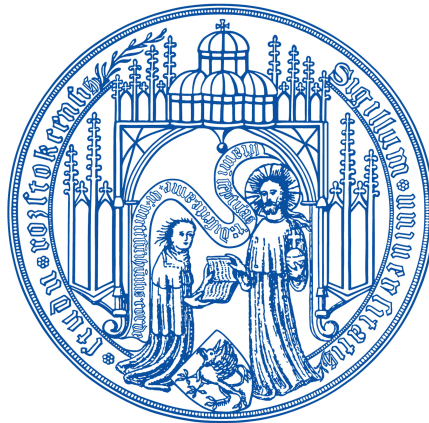

Towards a biological modelling tool recommending proper subnetworks

Master Thesis (revised)

Universität Rostock
Fakultät für Informatik und Elektrotechnik
Institut für Informatik



vorgelegt von:	Fabienne Lambusch
Erstgutachter:	Prof. Kurt Sandkuhl
Zweitgutachter:	Dr. Dagmar Waltemath
Betreuer:	Dipl.-Inf. Ron Henkel
Abgabedatum:	09. November 2016

Abstract

Modelling is an essential task in systems biology in order to describe complex biological systems and predict their behaviour. Computational models that represent biochemical reaction networks grow in size and numbers. Consequently, it is more likely that a given model portion has already been published and made accessible. Search engines provide means for finding and reusing the available models. Especially, large biological networks are often assembled from already existing networks. The amount and complexity of publicly available models makes it barely feasible to manually search and integrate proper networks for reuse in a currently developed model.

A tool that provides modellers with recommendations on how to extend their model will greatly benefit the users. Even though tools exist to support a user in searching, merging or combining models, there is still a high degree of manual effort required to perform reusing tasks. So far, there are no methods proposed in systems biology to obtain recommendations of suitable subnetworks based on the biological network under construction. Such an approach can bring together search, comparison, and integration of subnetworks.

The aim of this thesis is to develop methods that suggest the users suitable subnetworks for integration during modelling. To this end, techniques from the field of recommender systems are used, which aim to predict the users' interest in certain objects in order to filter and recommend the most suitable ones. Especially association rule mining is of particular relevance in this thesis. Its algorithms offer the opportunity to find patterns of joint appearance in a large set of items. For this purpose, biological networks are considered, which are represented as graphs and annotated with standardised ontology terms. Association rule mining then is applied with respect to structural and also to semantic similarity. For a partly modelled biological network the elements are found that may extend it. The obtained results form a solid basis for the development of a recommender system that facilitates the efficient reuse of networks and decreases the manual effort to find and integrate relevant structures.

Contents

1	Introduction	1
2	Background	3
2.1	Recommender systems	4
2.2	Recommender systems for business process modelling	7
2.3	Semantic similarity	8
2.4	Modelling and analysing biological networks	10
2.5	Further graph-based approaches	12
3	Requirements analysis	13
3.1	General requirements	13
3.2	Recommendation options	14
3.3	Survey	15
4	Model source, structure and storage	18
4.1	Data	18
4.2	Model storage and query language	21
5	Concept	25
5.1	Overview of prior findings	25
5.2	Incorporating semantic information	26
5.3	Strategy 1: based on a single entity	28
5.4	Strategy 2: one-to-one match of the whole network	29
5.5	Strategy 3: frequent patterns	29
5.6	Strategy 4: semantically similar models	30
5.7	Strategy 5: common subgraph	30
5.8	Strategy 6: entity list and edge estimation	31
6	Implementation	31
6.1	CellDesigner plug-in	31
6.2	Strategy 1: based on a single entity	32
6.3	Strategy 2.1: one-to-one match of the network structure	33
6.4	Strategy 2.2: one-to-one match of the network with annotations	34
6.5	Strategy 3: frequent patterns	34
6.6	Strategy 4: semantically similar models	35
6.7	Strategy 5: common subgraph	36
6.8	Strategy 6: entity list and edge estimation	36
7	Results	37
7.1	Strategy 1: based on a single entity	37
7.2	Strategy 2.1: one-to-one match of the network structure	38

7.3	Strategy 2.2: one-to-one match of the network with annotations	38
7.4	Strategy 4: semantically similar models	40
7.5	Strategy 6.2: entity list based on strategy 4 and edge estimation	41
8	Discussion	42
8.1	Findings	42
8.2	Conclusion	45
8.3	Future work	46
	Appendix A: Survey	52
	A.1 General results	52
	A.2 CellDesigner users only	60
	A.3 Cytoscape users only	68
	Appendix B: SBML models	76
	B.1 Exemplary SBML file	76
	Appendix C: Exemplary Cypher queries for strategies	77
	C.1 Strategy 1	77
	C.2 Strategy 2.1	77
	C.3 Strategy 2.2	78
	C.4 Strategy 4	78
	C.5 Strategy 6.2	79

List of Figures

1	Reaction graph of a SBML model	19
2	Detail of the reaction graph	20
3	A network model in the graph database	22
4	Semantic annotations in the graph database	23
5	Query example for a reaction with its participatory species adopted from (Lambusch, 2015)	24
6	Exemplary input structure for strategy 1	33
7	Exemplary input structure for strategy 2.1	34
8	Exemplary input structure for strategy 2.2	35
9	Exemplary query result for strategy 1	38
10	Exemplary query result for strategy 2.1	39
11	Exemplary query result for strategy 2.2	39
12	Exemplary query result for strategy 4	40
13	Exemplary query result for strategy 6.2	41

1 Introduction

An essential task in systems biology is to describe complex biological systems on an abstract level (Finkelstein et al., 2004). For this purpose, researchers model smaller components of such systems and reassemble them for a comprehensive understanding. Biochemical reaction networks are one of the occurring biological phenomena, which are driving for certain processes within a biological system (Hucka et al., 2003). Examples for corresponding biochemical reactions include phosphorylation and dephosphorylation. The Systems Biology Markup Language (SBML) is a standard notation to encode models that represent biochemical reaction networks (Hucka et al., 2003). In SBML, the participants of a reaction are called species. In the case of phosphorylation and dephosphorylation a participating species would be a protein kinase. Reactions become interconnected by sharing species and thus, built networks.

Computational models that represent biological networks become very complex as they grow in size (Randhawa et al., 2010). Models of large reaction networks are often constructed by combining already existing smaller subnetworks. As a consequence of the increasing complexity, it becomes more difficult to reuse existing networks and to create large models manually. A software tool that supports users in the efficient reuse of networks by providing them recommendations based on the currently developed model will greatly benefit the users. Nevertheless, so far there is no strategy proposed for such an assistance functionality in biological modelling tools.

There are various software systems for the graphical modelling of biological networks, for example CellDesigner (Funahashi et al., 2003), Cytoscape (Shannon et al., 2003), or SBGN-ED (Czauderna et al., 2010). Approaches exist to support the process of modelling, including functionality for finding the right annotations of elements (Krause et al., 2010) or verifying the structural and dynamical properties of models (Sadot et al., 2008; Heiner and Koch, 2004; Antoniotti et al., 2003). But the present support for reuse of models or parts of them still requires a lot of additional effort (Peng et al., 2013). In Randhawa et al. (2009), for example, biological components have to be defined specifically for the purpose of combination. Similarly, a precondition for merging models as proposed by Krause et al. (2010) or Randhawa et al. (2007) is the previous search for models and the selection of suitable ones. This demonstrates that a high manual effort is still needed to perform reusing and combining tasks. An approach for obtaining recommendations can bring together search, comparison, and the following integration of suitable networks in an automated manner. Such methods to obtain recommendations are missing in systems biology.

The aim of this thesis is to develop strategies to recommend suitable subnetworks that can be integrated in a user's biological network in order to expand it during modelling. The obtained results form a solid basis for the development of a recommender system that facilitates the efficient reuse of networks and decreases the manual effort to find and integrate relevant structures. Researchers could accelerate their modelling by effectively using their knowledge to combine biological data. There are several standard formats for modelling in systems biology. SBML, the Systems Biology Markup Language, is one of the best known standard formats (Li et al., 2010). It is used to encode models that represent biochemical reaction networks and is based on XML (Hucka et al., 2003). A

tool providing recommendations may help beginners and even advanced modellers to get into a so far unutilised graphical software and standard modelling formats by demonstrating exemplary structures that can easily be combined. By providing supportive, easily usable software tools for modelling in standard notations or rather with import and export functionality for several formats, one may encourage the more widespread use of these standards.

The research area of the so called recommender systems provides strategies to support users with recommendations (Ricci et al., 2011). By the versatile possibilities to distribute information and store large amounts of data quite an information overload can arise (O'Donovan and Smyth, 2005). If decisions shall be taken on the basis of this wealth of information, supporting a user is important (Melville and Sindhvani, 2011). The study of biological systems already produced vast amounts of data (Finkelstein et al., 2004). Furthermore, the number and complexity of publicly available computational models is still increasing. Recommender systems have emerged as valuable means to handle the information overload problem (O'Donovan and Smyth, 2005). These systems aim to predict the users' interest in certain objects in order to filter them from the unwieldy set of possible objects. As a result, they provide a specific user just the data of interest (Melville and Sindhvani, 2011). Recommender systems are used with increasing frequency and further research on their development is still required. Cases of application are primarily recommender systems in terms of product finders until now. Online shopping platforms suggest complementary products and instant video services recommend movies on the basis of similar users' ratings, just to mention two of the well-known examples. Another great contribution of recommender systems is the assistance in common working routines, especially for knowledge-intensive processes (Huber, 2015). In recent years, the effort has been made to utilise the concept of recommender systems for modelling procedures such as for business processes (see Section 2.2).

According to Butcher et al. (2004), "Systems biology aims to describe and to understand the operation of complex biological systems and ultimately to develop predictive models of human disease." Models of biochemical reaction networks can be represented as mathematical graphs (see Chapter 4). Such formal schemes facilitate the opportunity for computer-aided analysis (Finkelstein et al., 2004). Therefore, strategies for recommender systems may be well-applicable to suggest users existing biological networks during modelling. Mostly, data mining and machine learning algorithms lie at the heart of recommender systems (Ricci et al., 2011). One subclass of data mining is association rule mining, which is of particular relevance in this thesis. Its methods search for rules that predict the occurrence of an item based on the occurrences of other items, whereby the items' interactions must be available explicitly or implicitly. The graph representation of biological networks facilitates their transfer to the aforementioned scheme. Nodes in the network are the items and the edges represent explicit interactions between these entities. With association rule mining it becomes possible to find for a partly modelled biological network the elements, which often appear jointly with the given ones. The resulting subnetworks are the candidates to extend the user's model and thus, represent the recommendations. In this thesis, association rule mining is applied with respect to structural and also to semantic similarity. In Chapter 2, background information and related work are presented. A short survey was conducted that

indicates possible requirements for recommendations of biological networks. Chapter 3 deals with possible requirements and describes the content and results of the survey. As basis for recommendations, SBML-models from the public repository BioModels are used, which are stored in a graph database to utilise their graph structure. In Chapter 4 the used models and their representation are explained. To recommend suitable subnetworks, the structure and semantics of the user’s model under construction are considered. Six strategies are proposed to provide modellers of biological networks with recommendations. The concept is explained in Chapter 5 and exemplary implementations of several strategies are described in Chapter 6. The strategies are envisioned to be implemented as extension of an available modelling tool. All strategies are based on querying the graph database for extending structures. The proposed methods make use of biological models that are annotated with standardised ontology terms. These ontology terms and also the network structure are utilised to compare the networks and recommend extending subnetworks. The results of the used exemplary queries are described in Chapter 7. They demonstrate that methods of recommender systems are applicable for the domain of biological network modelling. In Chapter 8, the findings are discussed and conclusions are drawn.

It is revealed that it is feasible to recommend single network entities or semantically similar models adequately, whereas creating recommendations of complex fragments is quite difficult. The algorithms required to recommend larger subnetworks are often very complex and not scalable. Appropriate methods to recommend larger fragments shall be examined in future work. The proposed strategies can ease the reuse of existing biological networks by reducing the search space. Each strategy has its advantages and drawbacks. Therefore, providing users different strategies, such as the developed, allows them to decide for a strategy appropriate for their current modelling situation.

2 Background

This chapter addresses topics related to a recommendation functionality for biological networks. Section 2.1 describes the quite general term of recommender systems and possible subcategories. The focus of the section is especially on content-based recommendation approaches and a specific data mining category. Methods of data mining are relevant to infer additional information that can be used for recommendations (Ricci et al., 2011). In particular, a brief introduction in the category of association rule mining is given, which is used for strategies in Chapter 5.

Similarly to biological network models (see Chapter 4), business process models can be represented by graphs, for example to abstract from the several existing notations (Dijkman et al., 2011). The graph representations of business processes and biological processes can resemble each other in further aspects, such as in the use of labels or in linking to external semantic knowledge. As a result of these similarities, recommendation methods for business process models may be well adaptable to the biological models regarded in this thesis. Section 2.2 deals with insights over the recent years to utilise the concept of recommender systems for modelling business processes. To improve the reliability of recommendations, incorporating semantic information is important (Ricci

et al., 2011). Methods to compare data related to semantics applied for systems biology are considered in Section 2.4. For other domains semantic approaches are considered in Section 2.3. These approaches cover comparison of text, such as of graph labels, as well as comparison of semantic data from common schemes, namely ontology concepts. Biological papers are presented in Section 2.4. To address the wide range of topics related to recommender systems, the mentioned research topics here range from tool support for graphical modelling through merging, combining and searching to similarity scores of biological networks. Because of the graph representation of the models considered in this thesis, some approaches of the previous mentioned sections utilise graphs regarding a particular domain. Section 2.5 explains some general graph-based methods possibly relevant in the context of a recommendation engine.

2.1 Recommender systems

Recommender Systems aim to provide users with meaningful suggestions of items, which may be interesting or useful for these users (Melville and Sindhvani, 2011). Items to be recommended may be products to buy, such as music, films or news, whereby a recommender system usually concentrates on one selected type of item (Ricci et al., 2011). Generally, they predict a user’s rating for items and can then recommend, for example, a number of best matches (Adomavicius and Tuzhilin, 2005). At least, probable ratings are compared, if an exact prediction is not feasible or necessary (Ricci et al., 2011). The result can be represented as a ranked list of items based on the suitability referred to the user’s preferences.

Providing recommendations is a common feature especially in e-commerce (Fellmann et al., 2015). However, in the field of modelling, automated recommendations may support effective and efficient working of modellers. Ricci et al. (2011) states, that “Recommender systems have proven to be valuable means for online users to cope with the information overload [...]” by guiding users in a unwieldy set of possible options to the objects, which may be relevant.

Origins of recommender systems

The research area of recommender systems emerged in the 1990s and originated from the idea that people rely on recommendations provided by others with similar tastes (Ricci et al., 2011). Recommender systems comprise a wide spectrum of problems (Adomavicius and Tuzhilin, 2005). On the other hand, practical application is an important part of the research (Ricci et al., 2011). For example, recommendations play a key role for Amazon, YouTube, Netflix, Yahoo, Tripadvisor, Last.fm, and IMDb. As a consequence, there is a strong interest in the research and it is further increasing.

The existence of an ACM conference for recommender systems only since 2007 and the availability of Ricci et al. (2011) as a first comprehensive book for this topic are two examples, which demonstrate that this field of research is still relatively new and holds plenty of potential. Nevertheless, the theory for recommender systems utilises and intersects many established disciplines, particularly artificial intelligence with its subfields machine learning and data mining, human computer interaction, information

retrieval, marketing and many more. Data mining is essential for the strategies chosen in this thesis in order to generate from existing biological networks proper subnetworks for recommendation . Therefore, data mining is described in some more detail at the end of this section.

Data sources for recommendations

There are non-personalised recommendations like recommending the top ten of books, but typically recommendations are adjusted to the user's interests (Ricci et al., 2011). Therefore, recommender systems can use several information sources, including data about the users, the available items, or the interactions of the users with the system or specific items. Data about a user stems from explicit specification, as from ratings of items, or can be inferred from the aforementioned interactions of the user.

In this thesis, the focus is on predicting the best next network structures for a biological model based on a partly modelled network. Thus, covered approaches can be seen as personalised only in terms of inferences from the present network. Collecting several information about users for optimisation of results can be an approach to extend this work.

Classes of recommender systems

In the creation of a recommendation engine the properties of the available data as well as the domain of application are determinant for design decisions (Melville and Sindhvani, 2011). As mentioned before, recommendations are typically specialised in a certain item type within a system. Therefore, the architecture of the recommender system, such as the graphical user interface and the techniques to recommend items, are tailored to this specific item type (Ricci et al., 2011). Recommender systems can then be categorised referring to the chosen recommendation technique. There are two widely known main classes and a third that combines the both (Adomavicius and Tuzhilin, 2005). The most popular and most often implemented category is collaborative filtering (Ricci et al., 2011). Systems from these class provide a user recommendations in accordance with similar users by comparing the rating histories (Adomavicius and Tuzhilin, 2005). Thus, only historical interactions across users are analysed (Melville and Sindhvani, 2011). It is assumed that such an approach alone is not sufficient for modelling purposes, where recommendations should mainly based on the features of the models. The other category is content-based filtering. In contrast to collaborative filtering approaches, the recommendations of these systems are based on the attributes of the specific user profile together with the features of items, for example the genre for films (Ricci et al., 2011). As a consequence, recommended items are selected by matching the user's preferences for certain object features with the attributes of an item. Examples include the recommendation of programming books for software engineers or baby toys for a mother. The results can be ranked by the relevance according to the user profile. In this thesis no extra user profiles are considered. The user's preferences are inferred only from the network model under construction as a first step towards support in biological modelling. A user's history, for example, the clicking-behaviour as

in Hornung et al. (2008) is provisionally not considered. Information retrieval intersects significantly with content-based systems, because the recommendation techniques have their roots in this research area (Ricci et al., 2011). For both fields it is essential that users receive information relevant for their needs, wherefore an important component is a content-based search. Moreover, filtering and ranking are also crucial for both to succeed. The slight difference is that information retrieval typically uses global methods, whereas recommender systems often facilitate individual preferences of users and use a kind of interest or utility criteria.

Hybrid systems combine classes in order to enable the use of advantages of one approach to compensate the disadvantages of the other. Sometimes some further approaches are mentioned in literature, including demographic, community-based, or knowledge-based systems. Demographic techniques would use user data such as the age or country to filter recommendations. In community-based approaches, recommendations are constructed from the preferences of the user’s friends. This research is especially of interest with the rising popularity of social networks. Knowledge-based systems may be interesting for modelling purposes, because they utilise the domain context or rule knowledge to fit the needs of their users.

Association Rule Mining

As mentioned earlier in this section recommender systems utilises, among others, the research fields machine learning, data mining (Ricci et al., 2011). Algorithms of these two areas are core elements in many cases of recommender systems. They offer the opportunity to optimise the task performance by learning. Data mining generates new information from the plethora of available data, so that recommendations can be constructed on their basis.

One subclass of data mining is association rule mining, which is of particular relevance in this thesis. Ricci et al. (2011) states, that association rule mining “focuses on finding rules that will predict the occurrence of an item based on the occurrences of other items in a transaction”, whereby transactions must be available explicit or implicit. This matches the problem description of the thesis: given a partly modelled biological network in its graph representation (see Chapter 4), one wants to find the extending network elements, which often appear jointly with the given ones. An explicit notion of transactions is given by making use of the graph representation of the biological networks, where the edges represent the relations between nodes. Association rule mining can effectively detect patterns in data sets (Ricci et al., 2011). Its algorithms have proved to be even more accurate than a standard type of collaborative filtering algorithms.

Association rule mining investigates the possible sets of items together with their frequency of occurrence. The fraction of transactions, which contain an item set, is called the support of the set of items. Given the item sets X and Y , an association rule is an implication of the form $X \Rightarrow Y$. The fraction of transactions that contain both sets of items, is called the support of the association rule. Another important term is confidence. This describes how frequently items from Y occur in transactions containing the items of X . Given a threshold for the support and confidence value, the aim of as-

sociation rule mining is to detect all the rules that comply at least these thresholds for support and confidence. It is computationally very expensive to compute support and confidence in the same procedure for the rules. Thus, a two-step approach is typically performed. In the first phase all the sets of items with the desired support or more are generated. Afterwards, item sets with the highest possible confidence are built.

2.2 Recommender systems for business process modelling

A business process is a set of activities that produces an output valuable for a customer (Hammer and Champy, 1993). Most notations to model business processes are graph-based, so the models have nodes and relations between them (Dijkman et al., 2011). In this thesis, the graph representation of biological networks is used (see Chapter 4) and can resemble business processes in several aspects, such as in the use of labels or in linking to external semantic knowledge. Hence, approaches for recommendations in modelling business processes may be efficiently adaptable for biological networks.

The study of recommender systems in the context of business process models (BPMs) is relatively new. (Fellmann et al., 2015) states, that implementation strategies for a recommender system in process modelling were suggested, but were not yet exploited in commercial tools. The authors of this publication elaborate a requirements catalogue for the creation of recommender systems in the field of business process modelling. The described requirements served as an impulse to start a short survey in the context of systems biology for this thesis. Chapter 3 deals with possible requirements in the context of biological networks and the structure, as well as results of the survey.

The authors of Wieloch et al. (2011) develop a concept for semantic-based recommendations in the context of BPMs. Their methods search for fragments that may semantically follow an existing model fragment. Therefore, the context and annotations of elements are considered. The user has to explicitly choose a fragment for extension and one of five strategies. The user can select a resulting fragment from a list, which then substitutes the prior chosen fragment.

Smirnov et al. (2009) utilises association rule mining (see Section 2.1) to generate recommendations, but only focuses on activities in business processes. By using an interpretation function for activity labels, actions are inferred from the activities. The developed method identifies sets of actions that often appear jointly. Afterwards, the order of activities is determined for integration in the existing BPM.

The authors of Li et al. (2014) also use a kind of association rule mining for recommendations. They extract occurring subgraphs in a repository of business process graphs, whereby only the confidence and not the support of subgraphs is considered. Preprocessing of the repository’s models is required to facilitate the extraction. This means all the BP graphs are remodelled to be represented uniformly. Fragments are recommended as a ranked list by computing the distance between the user’s process model and the extracted patterns.

The recommendation method in Bobek et al. (2013) utilises the concept of Bayesian networks, which are acyclic graphs representing random variables and their dependencies. A business process graph can be transformed to a Bayesian network by modelling the nodes as random variables and the edges as the dependencies. The network needs

to be trained, for which an expectation maximisation algorithm is used. The recommendations are process-fragments as a ranked list, starting with the one that is most probable to be a missing part.

Hornung et al. (2008) describe support for modellers by a search interface and recommendations of process fragments. The user can choose the whole model or some elements for the search or recommendation. From the selected elements all labels are used to create tags (see Section 2.3) and a search query consists of the concatenation of these tags. The ranking of the results, for the search interface and recommendation, is the weighted combination of scores for the tag-based search, the syntactic correctness of the fragment and the frequency with which users have chosen the fragment in the past. A user can change and save fragments, by which they are included in future searches. Dijkman et al. (2011) consider the following problem: given a model or rather model fragment, the most similar models within a BP repository shall be found. Therefore, the authors develop three similarity measures for BPMs involving different information sources. These sources range from only considering element properties through the regard of the elements' topology to examining the behaviour. The search results for each measure can be ranked by the degree of the respective similarity. The first method computes an optimal matching between the models' process nodes by using their labels and attributes. The overall score stems from the inclusion of the total number of nodes in the compared models. For the structure-based search a graph edit distance is used utilising the node labels. The measure that approximates behavioural similarity includes label comparison and causal relationships.

2.3 Semantic similarity

Many publications that consider similarity of biological models emphasise the impact of semantic information (Thavappiragasam et al., 2014; Henkel et al., 2010; Schulz et al., 2011; Alm et al., 2015; Pesquita et al., 2009). Biochemical networks are often compared by use of their labels or semantic annotations (Schulz et al., 2011). In this section similar approaches from other domains are described, whereas Section 2.4 considers systems biology approaches.

Comparison of labels

The comparison of labels is common in business process modelling (Hornung et al., 2008; Dijkman et al., 2011; Smirnov et al., 2009; Wieloch et al., 2011). One option to compare character strings pairwise is the use of a string edit distance, where the number of insertions, deletions or substitutions determines the similarity degree (Dijkman et al., 2011; Maedche and Staab, 2002). A label preprocessing, such as eliminating spaces or underscores, may be necessary for a more uniform representation (Dijkman et al., 2011). Nevertheless, string edit distance can be misleading, if words are syntactical very close (Maedche and Staab, 2002). The words “power” and “tower” may have a high match, even though their meaning is not very close.

A further problem is the ambiguity of natural language (Ricci et al., 2011). String matching will typically fail, if a word has several meanings and furthermore, if different

words have the same meaning. For the latter case, a possible solution is to compare labels including the synonyms of all the words they consist of. The authors of Dijkman et al. (2011) propose to collect the synonyms for labels from a dictionary and to perform a pairwise label matching by the comparison of their synonym sets. The method in Hornung et al. (2008) directly compares networks by label synonyms instead of starting from node comparison. All labels from a network are extracted and the most important ones are selected by use of the “term and document frequency measure”. The network obtains tags, which correspond to all the synonyms of the selected labels. Comparing networks then means to compare their related sets of tags.

Comparison of ontology concepts

While labels can be ambiguous (Smirnov et al., 2009), ontologies formalise and conceptualise an application domain by providing a source of defined terms reflecting domain specific knowledge (Ricci et al., 2011; Maedche and Staab, 2002). Ontologies semantically describe items and represent their relationships (Ricci et al., 2011). Words can then be associated with concepts from an ontology (Li et al., 2003). Ontologies are important to measure the similarity between words, because often they are compared by means of the associated concepts. Annotations linking biological entities to terms from an ontology facilitate precise similarity measures (Pesquita et al., 2009).

A frequently cited publication that is significant for an ontology-based strategy in this thesis is Li et al. (2003). The authors develop a semantic similarity measure for concepts within a tree-like lexical taxonomy. Several strategies based on different information sources are examined and their combination possibilities are evaluated. It is mentioned that a similarity may be asymmetric, but experiments showed a deviation of maximal 5 percent on average. As a consequence, the authors do not consider asymmetry for their similarity measure. Three strategies are explained: the usage of the shortest path length, the depth, and the information content of the concepts.

In the computation of the path length between concepts three cases have to be addressed. If the concepts are the same, the path length is set to 0. If they are not the same, but share words in the hierarchy, then the length is set to 1. If the the concepts are not the same, nor share words, the path length is counted exactly. The depth of concepts is important, because terms in deeper levels of the hierarchy are semantically more concrete, they may closer resemble one another. Concept depth is computed by counting the levels from the hierarchy root to the concept. A value for the information content comes from the occurrence probability of an instance. The authors’ evaluation shows that considering the information content does not enhance their similarity measure. As a result of the evaluation, a non-linear combination of measures for depth and path length is suggested.

Blanchard et al. (2005) describes eight different similarity measures for ontology concepts. Among others, similar strategies as the depth, path length, and information content are mentioned. Depth and path length are also considered for annotation of BPMs (Wieloch et al., 2011). The authors use these two strategies similar as mentioned previously, but in the context of business function comparison.

2.4 Modelling and analysing biological networks

The recommendation strategies resulting from this thesis are envisioned to constitute an improvement within the modelling of biological networks. Therefore, a prototype should be implemented as an extension of an available biological modelling tool. The first part of this section describes two well-known biological modelling tools. The next part considers the combining of biochemical networks. Finally, similarity measures are described, which are important to search and rank biological networks.

Graphical software tools

There are various software tools to model biological networks graphically. One of the most commonly used tools is Cytoscape (Wang et al., 2015). It provides users with opportunities to layout and analyse networks (Shannon et al., 2003). Further data can be integrated from external databases, such as functional annotations. The core can be extended with numerous plug-ins¹ that support, among others, graph analysis.

Another software tool is CellDesigner, which facilitates a graphical representation of networks and is SBML-compliant (Funahashi et al., 2003). Users can for example search, import, edit and simulate models. CellDesigner can be extended by simply writing a Java program and adding it as plug-in².

Merging and combining networks

Despite the availability of modelling and analysis software, reusing parts of models requires still a lot of additional effort (Peng et al., 2013). Randhawa et al. (2007) state that there is a lack of support for combining or composing models. In their paper they describe the merging of two or more models as well as the composition of models from smaller submodels by facilitating language additions for SBML. Krause et al. (2010) describe a software for merging SBML-models, where elements are matched by use of their semantic annotations. The user can control the merging process, e.g. resolving conflicts, or can execute it automatically.

The editor proposed by Randhawa et al. (2009) lets a user explicitly define biological components for the purpose of combination as a kind of modularisation. The important parts of such a module are then the input and output port, whereas a user do not have to understand the details of the module. For their purpose they suggest adding language features to SBML. Methods like the above mentioned for combination of models or submodels are relatively close to the modelling support considered in this thesis, but lack the automated search, ranking and especially the recommendation of model parts that are relevant in the user's current modelling context.

Model search and similarity measures

As mentioned in Section 2.1, filtering and ranking are important parts of a recommender system (Ricci et al., 2011). Model search and ranking for systems biology is considered

¹<http://apps.cytoscape.org/>

²<http://www.celldesigner.org/plugins.html>

in Henkel et al. (2010) and Schulz et al. (2011). The authors of Henkel et al. (2010) present a concept to retrieve and rank biological models mainly by their semantic annotations. The keyword-based model retrieval supports users in filtering possible model candidates and the relevance ranking based on the user’s query is helpful for decision making. An implementation of the proposed methods was available as part of the search engine for the model repository BioModels Database.

Schulz et al. (2011) describes an approach to find to a given query model all the models that share many semantic concepts ranked by the similarity score. For this purpose as well as for clustering and alignment, the authors study similarity measures based on semantic annotations and suitable for the comparison of biological models. It is possible that annotations link an entity to several semantic resources. Therefore, different web resources are merged into one comprehensive ontology in this publication. This eases the computation of similarity by mapping equivalent concepts or adding relations between terms of different ontologies. Two approaches for similarity computation are then proposed. The first approach creates a vector for each model where the entries specify whether each biological concept is contained in the model. The similarity can be computed by typical vector functions, such as the cosine coefficient. The second approach starts by pairwise comparison of model elements and computes on this basis an entire, possibly weighted similarity score. The comparison of biological concepts is grounded on the findings in Li et al. (2003), which are described in Section 2.3. Thus, they consider the shortest path between concepts and depth in the ontology graph, but add examples for biology. The similarity measures were evaluated with models from the repository BioModels Database. Similarly to Li et al. (2003), the authors ascertain the sufficiency of path length and depth in ontologies, whereas incorporating the information content can not increase the accuracy of the similarity measure. Thereby, the applicability of the measures proposed by Li et al. (2003) for biological models is approved. Their approach is considered in Section 5.2.

Pesquita et al. (2009) examines several similarity measures based on ontologies, but its focus is on the applicability for an ontology named the Gene Ontology. The similarity measure for biological models in Thavappiragasam et al. (2014) combines the comparison of element labels with the comparison of annotated resource identifiers linking entities to semantic definitions. An implementation is available as the tool SBMLcompare, which computes the similarity between SBML-models. Label texts may be ambiguous. Therefore, label similarity is only used if no semantic annotations are available. The semantic information of elements is compared by simply examining the equivalence of the linked resource identifiers, whereas the label similarity is computed by a string edit distance. The comparison of reactions requires additional effort, because all participating species must be matched. Thus, the authors suggest to check the list of the models’ components first.

As the above-mentioned publications demonstrate, the impact of semantic information in systems biology is often emphasised for similarity of biological processes, but structural information such as relations between reactions and species may also be very important. The incorporation of structural information is regarded in Yang and Sze (2007), where graph matching for biological networks is considered. Given a graph of interest for a biologist, the subgraph with the highest similarity score in another

graph should be found. The similarity score is used to output the top-k results. The software GraphMatch is the respective implementation, but for incorporating biological knowledge like semantic annotations, an additional tool must extend it. Pržulj (2007) proposes a pure structure-based approach for the comparison of biological networks. It considers the distributions of nodes in the graph, “For example, we count how many nodes touch one triangle (i.e. graphlet G2), how many nodes touch two triangles, how many nodes touch three triangles, etc.”

The next section describes more general graph-based approaches for structural similarity.

2.5 Further graph-based approaches

In the former sections some graph-based approaches were already described in the context of specific domains. In this section more general methods for graph analysis are considered. Bunke and Shearer (1998) elucidates an error-tolerant graph matching metric. This is based on the maximal common subgraph, which is generated by regarding subgraph isomorphism. The authors state that this similarity metric is a graph edit distance with a certain cost function. In Bunke et al. (2002) two exact algorithms for generating the maximum common subgraph are compared. They are evaluated on randomly connected graphs.

In this thesis the following problem is considered: given a partially modelled network, the possible following subnetworks should be recommended. In Kim and Leskovec (2011) the authors consider a related problem, in which the data about a network are incomplete and all the missing nodes or edges should be inferred. The result is the algorithm KronEM and its implementation, which only uses structural and no additional information. For that, the probability distribution for the missing network part is considered, particularly by maximising the probability to find the optimum. The missing parts are estimated and the parameters are inferred in several, alternately iterations based on a user defined source of complete network examples. A precondition for this algorithm is the knowledge of the amount of data that is missing. The authors state that their algorithm is applicable, among others, for networks from the systems biology. Further, they declare that the algorithm is scalable for thousand of nodes and applicable even for networks with about 45 percent missing data.

3 Requirements analysis

Recommendation strategies have so far not been proposed for modelling biological networks. This chapter deals with possible requirements that arise from the architecture of recommender systems in general and in the case of modelling. No claim is made to completeness. Instead, a selection of requirements is presented that may be particularly important to assist modellers of biological networks in their modelling task. These selected requirements constitute a solid foundation to choose strategies appropriate for this thesis.

The first section describes more general requirements mainly referring to long-term usability of a recommender engine and the acceptance of modelling recommendations by the users. Section 3.2 deals with concrete options within the design of a recommendation engine for modelling. These are of special interest to choose appropriate recommendation strategies. Therefore, a short survey was conducted on the basis of possible recommendation variants. The considerations for the survey and its results are presented in Section 3.3.

3.1 General requirements

Typical general requirements include, among others, scalability, usability and extensibility (Fellmann et al., 2015). These are not specific to recommender systems, but may be decisive to develop appropriate recommendation strategies. Considering the increasing amount and complexity of biological models (Courtot et al., 2011; Randhawa et al., 2010), especially the scalability of the system is important (Ricci et al., 2011). In view of the abundance of biological tools for different purposes (see Section 2.4), a recommender system should be compatible to available tools and languages (Fellmann et al., 2015). Recommendations should utilise or support at least one of the common standard formats for biochemical reaction networks or should even be language independent. One way to achieve this is the integration of a recommender system into one of the existing, well-known modelling tools for biological networks (for examples see Section 2.4). To fit different recommendation needs, it should be possible to choose between different recommendation strategies. Extensibility can be achieved by providing common interfaces (Wieloch et al., 2011). Implementing different strategies with the same interface allows for addition of further recommendation approaches. Moreover, a plug-in interface facilitates enriching the system with recommendation independent functionality. Another important requirement is the support of evolutionary changes in the recommendation source (Fellmann et al., 2015). An option to integrate new data in an operational recommender system is to incorporate learning algorithms (Ricci et al., 2011; Melville and Sindhvani, 2011).

Biological models can be built at different levels of detail (Finkelstein et al., 2004). Thus, recommendations should comply with the user’s modelled degree of abstraction and ensure a high semantic quality (Fellmann et al., 2015). The system should only recommend meaningful items that are not yet present in the currently developed model. Recommendations should provide users with diverse items (Ricci et al., 2011). This can be problematic, because the system developer has to trade off the diversity of recom-

mended items against the accuracy. An improvement may be achieved by empowering users to give the system feedback or to intervene, for example with buttons “I do not need this” and “show me more of this”.

Similarly to results of search approaches as in (Henkel et al., 2010), it may be difficult for users to judge the suitability of listed recommendations. Therefore, it is very important to help users make good decisions, while spending less time (Ricci et al., 2011). On the one hand, this effectiveness and efficiency can be reached by supporting the user with diverse information about the items, for example with images. On the other hand, the user should be able to adjust the displayed results, for example, with the help of filters (Fellmann et al., 2015). In summary, following requirements should be considered:

- Scalability of the recommendation algorithms
- Compatibility to common modelling languages for biochemical reaction networks
- Integration in a well-known modelling tool for biological networks
- Opportunities for recommendation dependent and independent extension
- Support of evolutionary changes in the recommendation source
- Provision of meaningful recommendations that have the right abstraction level and ensure a high quality
- Appropriate trade-off between recommendation accuracy and diversity
- Support of features for effective and efficient modelling

The next section considers several options for the implementation, such as the recommendation content, filters and displayed information about items.

3.2 Recommendation options

This section deals with design options for a recommender system in the context of modelling biological networks. Especially, the desired information content is important to choose appropriate recommendation strategies. A short survey was conducted on the basis of the recommendation variants, which is described in the next section.

If the user invokes the recommender system, it could suggest extensions for the network at any place or only for the latest added element. Alternatively, the user has to explicitly select elements for extension. In the latter case the selection can comprise a single element, a set of elements or a connected fragment (Wieloch et al., 2011). As basis for recommendations, the system should use at least the chosen elements, but even for a single selected entity the basis for the computation can be only the single element or the whole network. The recommended items are in turn single elements or connected element groups. Another option would be the recommendation of similar models (Fellmann et al., 2015). This is only considered as an additional feature in this thesis, because it contradicts the idea of directly extending a network.

The recommendation results can be represented to the user in a graphical view or table-based (Hornung et al., 2008) and sorted by a ranking (Ricci et al., 2011).

The system should support the users in identifying suitable items fast. Therefore, providing background information related to the items' properties and the quality of the recommendations is necessary. These information can include images of the graph structures, verbose descriptions of the fragments, the labels of the elements, a preview of related model parts, the frequency of the fragment within the knowledge base (Hornung et al., 2008), or the matching score (Li et al., 2014). In the context of biological models, displayed information can also include the ontology concepts of the elements and information related to the model, such as the publication or the authors. Furthermore, the system can provide additional explanations, why the item is recommended, what the differences to other recommendations are and to which extend it fulfils the goals (Ricci et al., 2011). The recommender system should provide filters to refine the results. The user could limit the amount of recommendations or the size of the recommended fragments. Further constraints could be set related to the afore-mentioned information, such as ontology types. In contrast, a list of results can be displayed first, where the user then can specify preferences for certain recommendations and the system refines the results based on the preferences.

3.3 Survey

A short survey was conducted to get an impression of what researchers, who model biological networks graphically, may expect from a recommendation functionality. To this end, research questions were drawn up based on the aforementioned options for outcome and representation of recommendations. The main objectives of the survey are covered by these research questions:

- Do modellers in system biology perceive it as helpful to get recommendations on how the model can be expanded?
- What should be recommended?
- What should be the data source for recommendations?
- Which information is necessary to choose suitable recommendations?

Each research question is associated with several questions in the survey. Due to the limited time frame of the thesis, the decision was taken to restrict the survey mainly to closed questions. The participants could add other answers in some questions and give comments at the end. In two question a multi-staged scale was used to provide participants the opportunity to expressing their opinion differentiated. Therefore, a scale between the numbers one and five was chosen, where only the endpoints are named verbally. The advantage here is that the interval scaling is equal in the distance. The scale ranges from unimportant on the left to important on the right. The odd number allows a neutral opinion. Additionally, some kind of escape check box was added for participants, who can not take a decision or do not want to take one. The survey includes 15 questions in total. Appendix A.1 shows all the questions and the general results of the survey. The recommendation strategies resulting from this thesis are envisioned to constitute an improvement within the modelling of biological networks. Therefore, a

prototype should be implemented as an extension of an available biological modelling tool. The survey results give an indication that the tools CellDesigner and Cytoscape are used preferentially to model biological networks graphically. Thus, the survey results were additionally filtered for both user groups. The answers of CellDesigner users are shown in Appendix A.2 and the answers of Cytoscape users in Appendix A.3.

The survey was available online via the website umfrageonline.com. The request for participation was addressed to modellers of biological networks and sent via seven mailing lists referring to BioModelsNet, CellML, SBML, BioPAX, Cytoscape, COMBINE and SBGN. It has to be emphasised that the survey was only conducted to get an impression of possible requirements and the results might not be representative.

Survey results

There were 47 participants in total. It is assumed that modellers have to use graphical tools so that they can answer the questions. Therefore, 16 of the participants dropped out of the survey, because one person does not find graphical tools helpful and 15 persons do not use such a tool. The majority of the questions was optional. The remaining 31 participants did not answer all the other question. The participation in each question is between 30 and 21 persons.

The most commonly tools used by participants are CellDesigner (37,9 percent), Cytoscape (37,9 percent) and SBGN-ED (20,7 percent). There are 17,2 percent of the participants, who use CellDesigner and Cytoscape. Respectively, 20,7 percent of the participants use only one of the both. The reuse of models or parts of them is seen as helpful by 100 percent of 30 participants and only some Cytoscape users think that a graphical modelling tool can not support them in the reuse (6,9 percent).

89,3 percent of the participants would find it helpful to get recommendations of models similar to their currently developed one. Thereby, the result for CellDesigner and Cytoscape users is 100 percent. To get recommendations of subnetworks is seen as helpful by 85,7 percent of the participants, whereby all CellDesigner users and only 90,9 percent of the Cytoscape users see it as helpful. The majority of participants would prefer recommendations of similar models (59,3 percent). In contrast, more than the half of the CellDesigner user would prefer subnetworks as recommendations (63,6 percent). In the case of recommended subnetworks 61,5 percent of the participants prefer structures with more than one element instead of single elements.

Most participants find a repository like BioModels Database as data source for recommendations helpful (92,6 percent), for CellDesigner users it is even 100 percent. The proportion of participants, who find a thematic subset of models or a customised set helpful as data source, is smaller (66,7 and 76,0 percent). Only 54,4 percent of the CellDesigner users find a custom set helpful. In contrast, the result for Cytoscape users is 70 percent. Regarding the information needed when selecting appropriate recommendation, images of the recommended structures seem to be important (57,69 percent) and also the names of the elements (50 percent). The importance of obtaining other information is controversial.

To limit the number of recommendations it is possible to implement filters. The majority of participants considers it helpful to have filters for recommendations (84,6 percent).

The importance of several options for filters is rated relatively varied. The restriction of the recommendations' maximum number tend to be important, whereas the minimum size of recommended structures tend to be less important. The importance of other filter options is particularly controversial.

Survey conclusion

It has to be emphasised that the survey results are not statistically proven, but give a first impression of possible requirements. It is surprising that all but one participants consider graphical tools to be helpful for the creation of biological networks, but nonetheless 32,6 percent of them do not use such tools. Thus, an easily usable tool providing recommendations may help beginners and even advanced modellers to get into graphical tools by demonstrating exemplary structures that can easily be combined.

As mentioned before, a future implementation of the proposed recommendation strategies should extend an available biological modelling tool. The survey results indicate that the tools CellDesigner and Cytoscape are used preferentially and could constitute a good basis for extension. Only based on the answers of the survey participants, CellDesigner users might be the user group to which recommendations are more attractive. Some of the participating Cytoscape users do not even think that a graphical modelling tool can support them in model reuse. All of the participating CellDesigner users see it as helpful to get recommendations of subnetworks and more than the half of them would prefer these recommendations to the ones of similar models.

In general, the survey results give an indication that reuse of models is helpful in systems biology, which corresponds to statements in the literature (Henkel et al., 2010; Courtot et al., 2011; Randhawa et al., 2010). Furthermore, Randhawa et al. (2007) mentions the lack of tool support for combining or composing models. A recommender system for biological networks that facilitates efficient reuse and integration may be exactly what is missing. A whole repository seems to be an appropriate data source for recommendations. Therefore, in this thesis models are used as data source that come from the curated branch of BioModels and are stored in a graph database (see Chapter 4).

The survey results indicate that also similar models and single elements as recommendations might be helpful. As a consequence, it might be good to have either a recommendation strategy, which is flexible and can recommend structures from single elements to whole models, or to have a modularised system where different strategies can be chosen. The strategies resulting from this thesis are described in Chapter 5.

In the context of information needed to choose suitable recommendations, images of the structures might be of particular interest and maybe also the labels and ontology terms of the elements. Thereby, ontology terms are controversial, but in contrast to labels they are unambiguous (see Sections 2.3 and 2.4). Integrating ontology terms in a recommender system may support the more widespread use of them. Participants added additional lines for information, which may be of interest for selecting recommendations. This indicates that further information seem to be required. To examine possible options and their adequacy is beyond the scope of this thesis and left open for future work. Filters seem to be relevant and should extend the proposed approach

in the future. The feedback given by participants point out the problem of the severe restriction within the possible answers. A comprehensive survey with options to express opinions freely can be a further effort in the future.

4 Model source, structure and storage

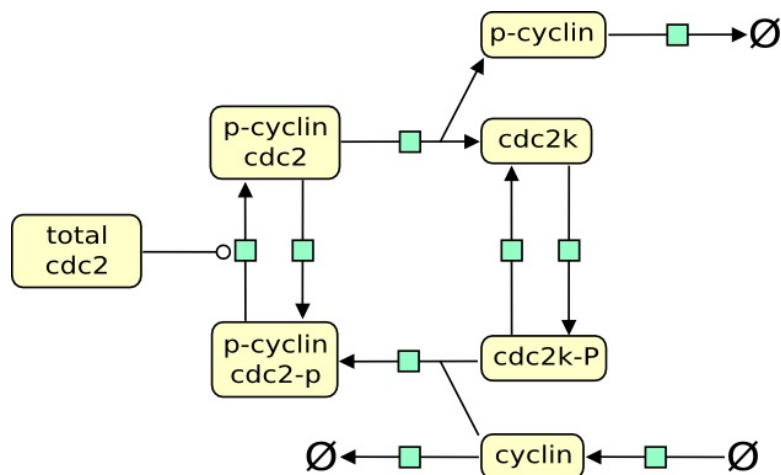
The aim of this thesis is to develop strategies to recommend researchers during modelling extending subnetworks for their biological networks. The basis for recommendations is the reuse of existing model networks. The survey results described in Section 3.3 indicate that a large public repository may be an appropriate data source for the reuse of biological models. For the strategies developed in this thesis, models from the repository BioModels³ in the standard format SBML are used. The first section provides more detailed information about the repository and SBML-models. The biological networks of the SBML-models can be represented as mathematical graphs. To utilise this graph structure, the models are stored in a graph database. Section 4.2 explains the model structure in the graph database and the use of the query language Cypher.

4.1 Data

The description of models by means of standard formats is very important in systems biology (Li et al., 2010; Stanford et al., 2015). The use of common standards facilitates the efficient exchange, reuse and comparison of biological models. Exchangeable standard formats allow different software tools to interpret and process the models (Henkel et al., 2012). There are different standard encodings in systems biology, for example, CellML (Lloyd et al., 2004) and NeuroML (Gleeson et al., 2010), but the most successful standard modelling language adopted by a wide range of software tools is SBML (Li et al., 2010). SBML is the abbreviation for the “Systems Biology Markup Language”, which is an XML-based format to encode models of biochemical reaction networks (Hucka et al., 2003). For the recommendation of suitable subnetworks, the focus in this thesis is only on models encoded in this well-known format. Biochemical reaction networks represent interactions within components of an organism, for example processes of the energy-metabolism of a living cell (Heiner and Koch, 2004). SBML encodes these interactions with their corresponding entities in a model, namely the reactions and species (Stanford et al., 2015). Hucka et al. (2003) defines a reaction as “some transformation, transport or binding process, typically a chemical reaction, that can change one or more chemical species”. These species can be entities such as ions or molecules and participate in a reaction. Phosphorylation and dephosphorylation are possible examples for reactions, whereby a participating species could be a protein kinase. A reaction with its participating species corresponds to a biochemical equation. Several reactions together built networks, if they share some species. These biological networks are commonly represented and treated as mathematical graphs (Yang and Sze, 2007). An example of a reaction network in its graph representation is shown in Figure 1.

³<http://www.ebi.ac.uk/biomodels-main/>

Figure 1: Reaction graph of a SBML model



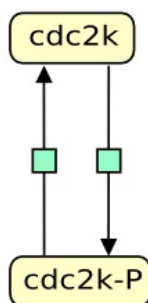
The figure shows the reaction graph of the model named “Tyson1991 - Cell Cycle 6 var” with the BioModels ID “BIOMD0000000005” exported from BioModels. The network is built from vertices for reactions (green) and for species (yellow), which are connected by edges representing the role of a species in a reaction as reactant (no arrowhead), product (standard arrowhead) or modifier (circle arrowhead). The empty set is encoded as a single species in the model, but occurs repeatedly in the figure.

The vertices of the graph correspond to the reactions and species of the model, whereas the edges correspond to the relations of these entities. Such reaction networks are bipartite graphs. This means that there are two disjoint sets of vertices, the reactions and species, and each edge connects a species vertex to a reaction vertex. Species can take a role in a reaction as reactant, product or modifier (Hucka et al., 2008). Modifiers do not actively take part in a reaction, but influence it. In a SBML-file, there is a list for both species and reactions and in each XML-block for a reaction all its corresponding reactants, products and modifiers are listed again separately (see Appendix B.1). However, in a reaction graph the species roles are typically illustrated by different arrowheads of the edges, as shown in Figure 1. The figure contains eight species and nine reactions. The species representing the empty set is counted once, because it is encoded as a single species in the SBML-file. It has a special role and thus, is shown in the figure multiple times. A species that is connected to a reaction by an edge without an arrowhead represents a reactant, whereas an edge with a circle at the end describes a modifier relation. Considering a reaction, an outgoing edge denotes the production of the connected species by the reaction.

The species labels in the reaction graph typically are the names of the species in the model. In the SBML-file, names for both species and reactions can be defined (Hucka et al., 2008). The name attribute is intended to provide a meaningful human-readable designator. The *name* is an optional attribute, whereas the *id* is a mandatory attribute and unique within all component identifiers of a model. Thus, the identifier can be

used to refer to a component, such as the reactions refer to their participating species. Humans do often disagree in the naming, which is why the identifiers and names in SBML-models are rather unrestricted to allow the users a high degree of freedom in choosing these attribute values. Furthermore, regulating the names by forcing the same names for common entities would require changes in the SBML-specification for every change in the list of predefined names. The absence of a standardized set of names or identifiers makes it difficult, especially for software tools, to determine the models' semantics reliably. Inferring the semantics for complex models may be infeasible even by means of a detailed analysis of the model components. Therefore, SBML provides the opportunity to add annotations to the elements, which link the entities to external resources declaring the semantics unambiguously. More precisely, the annotations connect the entities to unique database identifiers or ontology terms (Krause et al., 2010). It is further possible to express different relationship types between SBML-elements and resources by using biological qualifiers (Hucka et al., 2008). Summarising, SBML-entities can have a biochemical annotation that consists of a qualifier specifying the logical relation of the element to a resource identified by an URI (Schulz et al., 2011). The annotations can provide different information ranging from the biological meaning of a model entity through their role in a process to the type of an entity (Rosenke and Waltemath, 2014). Figure 2 shows a subnetwork of the model considered before. In the following, annotations are described exemplarily for this subnetwork.

Figure 2: Detail of the reaction graph



The figure shows the cyclic structure at the right of Figure 1

The annotations are not shown in the reaction graph, but are contained in the SBML-file (see Appendix B.1). There, *cdc2k* is linked to a resource in the protein database UniProt that identifies it as a *version of a cyclin-dependent kinase*. *Cdc2k* is further a reactant in a reaction, which can be identified as a *protein phosphorylation* by the reference to the Gene Ontology (GO). This reaction produces in turn a *version of a cyclin-dependent kinase* that is now phosphorylated, the species *cdc2k-P*. The shown structure is cyclic, because it also shows the possible *protein dephosphorylation* of *cdc2k-P* to *cdc2k*.

An ontology comprises several curated concepts that can be interconnected by various relationship types in a form, such as (“glucose”, “is_a”, “sugar”) (Schulz et al., 2011). The Systems Biology Ontology (SBO) has a special role in the SBML format that includes an extra *sboTerm* attribute (Hucka et al., 2008). SBO-terms are used to precisely

identify most SBML-elements. The use of the `sboTerm` attribute is optional and less flexible than the annotation option, but should as far as possible always be used for SBO-terms.

There are more elements in SBML-models than the above mentioned, such as stoichiometry and rate laws Hucka et al. (2003). However, the focus of this thesis is only on the reaction graph, whereby for the reactions and species the name, id and annotations are considered. Appendix B.1 shows an example for a SBML-structure, where the focus is on elements regarded in this thesis. Models in standard formats are distributed via public databases, such as BioModels Database (Li et al., 2010) or JWS Online (Snoep and Olivier, 2003). By means of the already established model databases, standard formats, and semantic knowledge bases it is fostered to search for model parts, to compare, and to reuse them, ultimately by automatic analysis (Li et al., 2010; Schulz et al., 2012; Henkel et al., 2016; Schulz et al., 2011).

The survey results described in Section 3.3 indicate that a large public repository may be an appropriate data source for the reuse of biological models. Therefore, models from the repository BioModels⁴ in the standard format SBML are used as a basis for the strategies of this thesis to recommend researchers proper subnetworks for network extension. BioModels is precisely intended for free distribution of models in standardised formats, which are curated and annotated according to defined standards (Li et al., 2010). Furthermore, BioModels is growing in popularity. According to Schulz et al. (2011), it is “the largest public collection of curated SBML models”. In this thesis, SBML-models from the curated branch of BioModels Database release 30⁵ are used. To utilise their graph structure, they are stored in a graph database. The next section provides more detailed information about this storage and the possibilities to query the graph database.

4.2 Model storage and query language

As the source for the recommendation of suitable subnetworks, the strategies developed in this thesis reuse existing model networks. To utilise their graph structure, the models are stored in a graph database. This section explains the model structure in the graph database and the use of the query language Cypher.

Biological networks in the graph database MaSyMoS

Graph databases are non-relational databases (Partner and Vukotic, 2012). To store and query biological models, the open-source graph database Neo4j is used (Hunger, 2014). Models are extracted from BioModels Database release 30 and stored in a Neo4j database called MaSyMoS⁶. MaSyMoS contains 1424 models, whereof 590 models are curated and in the standard format SBML. Only these curated SBML-models are considered in this thesis. The graph representation of these network models provides opportunities for a formal analysis (Finkelstein et al., 2004). Neo4J facilitates the easier

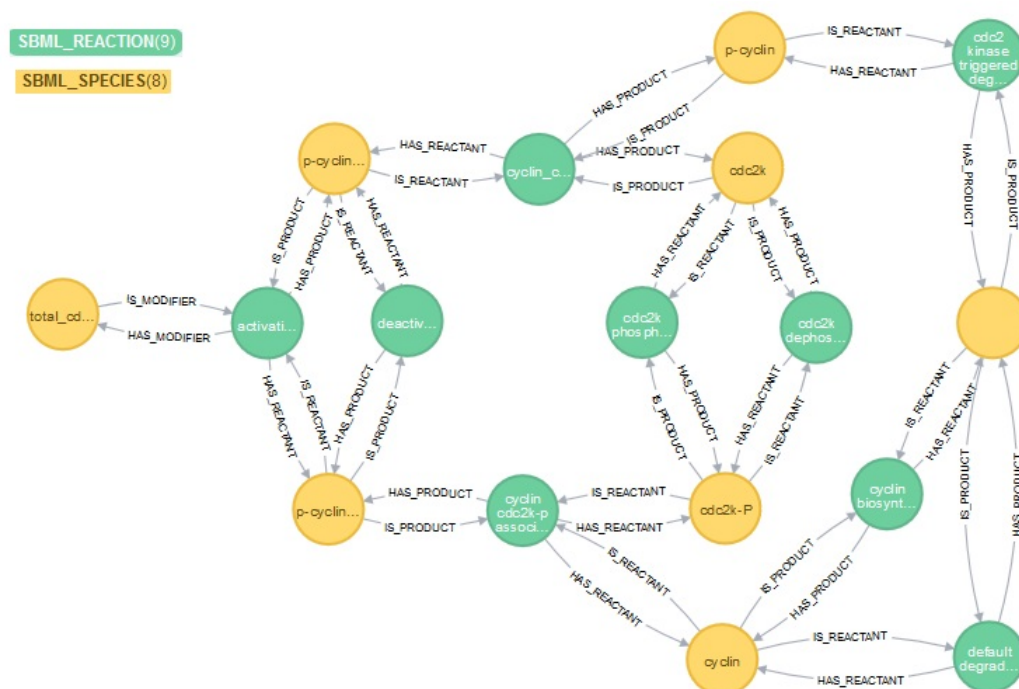
⁴<http://www.ebi.ac.uk/biomodels-main/>

⁵<ftp://ftp.ebi.ac.uk/pub/databases/biomodels/releases/2016-05-10/>

⁶<https://sems.uni-rostock.de/projects/masymos/>

use of well-founded graph theory algorithms (Partner and Vukotic, 2012). An example is that Neo4j monitors in a graph traversal the nodes already passed through with the result, that it visits each node just once.

Figure 3: A network model in the graph database



The figure shows the SBML-model named “Tyson1991 - Cell Cycle 6 var” with the BioModels ID “BIOMD0000000005” retrieved visually from MaSyMoS.

The network is built from vertices for reactions (green) and species (yellow), which are related by directed edges, one for each direction, representing a reactant, product, or modifier role of a species in a reaction.

The graph structure in Neo4J consists of vertices, which are related by directed, typed edges (Hunger, 2014). In MaSyMoS the biological network is built from vertices for reactions and species (Henkel et al., 2014). These are related by directed edges that represent the role of a species in a reaction as a reactant, product, or modifier. Therefore, there are always two edges between a reaction and a species representing that a reaction has a participating species and in turn that a species is participant in a reaction. All vertices have an internal identifier in the database and also a label that represents the node type, such as *SBML SPECIES* or *SBML REACTION* (see Figure 3). In a Neo4J

database, attribute-value-pairs can be added to the vertices and edges (Hunger, 2014). MaSyMoS assigns the additional properties *ID* and *NAME* to the vertices for reactions and species referring to the corresponding attributes in the SBML-file (see Appendix B.1) (Henkel et al., 2014). As one can see, there is one species node in Figure 3, which shows no name. This species represents the empty set that is shown multiple times in Figure 1. Models in MaSyMoS contain the empty set only as often as defined in the SBML-file.

For each semantic annotation in a model, there is a node of the type *ANNOTATION* in MaSyMoS that is related to the corresponding model entity. By means of an annotation node, a particular model entity is connected to *RESOURCE* nodes that each represent a particular term in the referenced bio-ontology. Every *RESOURCE* node has an attribute named *URI* in accordance with the particular resource URI in the SBML-file. The edge between the annotation node and the resource node corresponds to the defined biological qualifier of the model entity for the resource. Figure 4 shows an example for the visual depiction of annotations and resources in MaSyMoS corresponding to the exemplary SBML-file in Appendix B.1.

Figure 4: Semantic annotations in the graph database



The figure shows the annotations for the species *cdc2k* and the reaction *cdc2k phosphorylation* from the model displayed in Figure 3 and retrieved visually from MaSyMoS.

There is an annotation node (red) for the species and reaction, which links the entities by the biological qualifiers to the corresponding resources (grey)

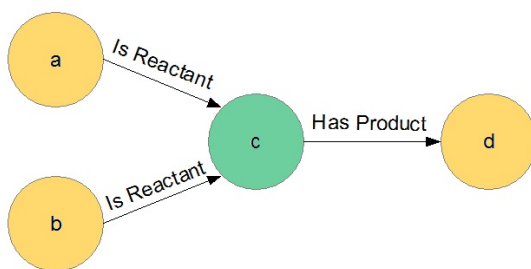
The figure shows the related resource nodes for the species named *cdc2k* and the reaction named *cdc2k phosphorylation*. For example, *cdc2k* is connected via an annotation node to a resource node, which has an URI value linking to the concept *cyclin-dependent kinase* in the protein database UniProt. The figure shows only the structure without the properties, but as one can see the edge type represents the biological qualifier *isVersionOf* as it is contained in the SBML-file (see Appendix B.1). The SBO-terms are represented in MaSyMoS in the same way as other resources, but have the edge type *HAS_SBOTERM* that connects the annotation node to the resource. The node for a SBO-resource has also a URI property and its value links to the particular SBO concept. In general, all concepts of an ontology with their corresponding relations are stored once

in MaSyMoS and can then be referred to by multiple entities from several models.

The query language Cypher

To query the Neo4J database MaSyMoS, one can use the language Cypher. It is a declarative language that is specialized in graph querying (Hunger, 2014). There are alternative query languages for Neo4J databases such as SPARQL or Gremlin, but Cypher is commonly used for Neo4J (Robinson et al., 2013). It allows the precise description of graphs and is quite easy to understand. One can intuitively describe a graph pattern that shall be matched with instances in the database. Figure 5 shows such a graph pattern that can be used to query the database with Cypher. It shows a reaction with its participatory species.

Figure 5: Query example for a reaction with its participatory species adopted from (Lambusch, 2015)



The figure shows a graph structure for a reaction (green) with its participatory species (yellow).

It can be used to query a Neo4J database intuitively with the language Cypher, similar to drawing the graph.

The figure shows the vertices *a* and *b* connected to a vertex *c* by the relation *Is_Reactant*, which is synonymous with two species *a* and *b* taking a role as reactant in a reaction *c* (Lambusch, 2015). The reaction *c* is in turn related to *d* by an edge of the type *Is_Product* representing that the reaction produces a species *d*. One can describe this graph pattern with Cypher similarly:

```
MATCH
(a)-[:Is_Reactant]->(c),
(b)-[:Is_Reactant]->(c)-[:Has_Product]->(d)
```

This defines a search for all database entries with exactly the given structure. A *RETURN*-clause must be specified to receive a desired result. To get the nodes of the matching reaction and species, one can add the line

```
RETURN a, b, c, d
```

at the end. This will return for each match the corresponding reaction with its connected three species. If the types of the nodes are known, it is possible to define a more

efficient query by adding the node type. For example one can refine (a) by using (*a:SBML_SPECIES*) instead. Furthermore, it is possible to define a specific starting point, for example, if only structures around a reaction with the name *phosphorylation* are searched. In this case, a *START*-clause can be added at the beginning of the Cypher query, such as the following:

```
START c=node:index-name(name='phosphorylation')
```

Further information about Cypher and possible clauses is provided by Robinson et al. (2013).

5 Concept

To introduce the developed strategies, an overview of prior chapters is given first. Section 5.2 describes the selected methods to incorporate semantics in the strategies. Afterwards, separate sections give detailed information about each developed strategy.

5.1 Overview of prior findings

Existing approaches to reuse biological network models or parts of them are limited only to support users in certain reuse steps, for example model search (see Section 2.4). Combining different approaches or tools requires additional effort and may be quite time-consuming. One of the existing tools that explicitly supports the creation of large networks by combining smaller ones, requires the users to define biological components beforehand specifically for the purpose of combination. Such approaches are not particularly automated and can not actively support the user with recommendations on how to extend their model. A comprehensive approach for recommendations of proper sub-networks during modelling and their consequential integration in the model is missing. The research area of recommender systems provides strategies to support users with recommendations (see Sections 2.1 and 2.2). Recommender systems guide users in a large set of options to items that may be relevant to the users' tasks and thus, support users in decision-making. In general, a recommender system is specialised in a specific type of items. The design, graphical user interface and recommendation techniques are precisely customised to this item type. The items considered in this thesis are biological networks in their graph representation. Furthermore, the strategies developed are personalised in the way that a user's preference is inferred from the currently modelled network. Long-term preferences are not considered in this thesis. By recommending networks based on the currently developed network model, it is not necessary to define components for combination beforehand. Instead, it becomes possible to utilise the large amount of already available models, because the tool can prefilter relevant networks according to the network under construction. The survey results (see Section 3.3) indicate that a whole repository may be an appropriate data source for recommendations. In this thesis, SBML-models from the curated branch of BioModels are used as data source, which are stored in the graph database MaSyMoS (see Chapter 4). The developed strategies compare the utility of subnetworks and output a ranked list of the subnetworks that may be the most suitable once for the extension. The basic idea

is to search for existing networks that are similar to the currently developed network with regard to the structure or semantic, and to recommend extending fragments from these similar models. In general, similarity measures for biological models (see Section 2.4) are only marginally useful for recommendation purposes, because they are designed to compare complete models, instead of considering a partly modelled network. This is especially problematic in the adaptation of structural similarity measures, which often rely on the size of the networks. Therefore, methods are required that focus on subsets of models. For example, different kinds of graph matching are already considered for biological networks. Given a graph, these methods mostly search for similar or isomorphic subgraphs within another graph. There are mainly two problems to adapt this for a recommender system. On the one hand the time complexity of these algorithms is problematic and on the other hand it is not clear, how to generate recommendations based on the matching structure. Most likely, all surrounding elements for the match have to be examined again.

Data mining is often used in recommender systems to gain new knowledge. It is also possible to use the information provided by data mining techniques to construct recommendations. Association rule mining is of particular interest for this thesis (see Section 2.1 for general information). With association rule mining it is possible to find the network elements that appear frequently together in a set of models. The elements, which are most likely to occur with the once of the partly modelled biological network, are assumed to extend it appropriately. Association rule mining is applicable for recommendations of graph structures during modelling, which is demonstrated by existing approaches in the domain of business process modelling (see Section 2.2). The survey results (see Section 3.3) indicate that several kinds of recommendations might be helpful for the users, ranging from single elements, through larger fragments, to similar models. Furthermore, incorporating semantics in the recommendation strategy is important to regard the semantic context and the right level of granularity in accordance to the model under construction. As described in Sections 2.3 and 2.4 there are several options to calculate the semantic similarity for biological networks. With respect to structure and semantics, a variety of options exist to support the users with recommendations during modelling. Therefore, several strategies are developed that are envisioned to be available concurrently. This allows users to choose the strategy, which best suits to their current task. The next section describes the selected methods to incorporate semantics in the strategies. Afterwards, a diversity of recommendation strategies is presented.

5.2 Incorporating semantic information

To regard domain knowledge and provide more reliable recommendations it is important to consider the semantics of models. In particular, the comparison of single network elements, namely the vertices and edges, becomes possible by regarding their semantic, whereas structural similarity measures are only useful to compare fragments. The strategies described in the next sections involve, in one way or another, semantic comparison of network elements. Therefore, this section describes the selected methods for the calculation of semantic similarity between biological networks by referring to findings of prior chapters, mainly Sections 2.3 and 2.4.

In this thesis, edges of a network are always compared by checking the equivalence of their type, which is all the semantic information provided by MaSyMoS (see Section 4.2). To compare vertices, three approaches are chosen and a fourth is combined from the others. The first approach is to simply compare the vertex types as in the case of edges. On the one hand, recommending subnetworks with so little information grants the users freedom to choose the biological meaning and granularity on their own, whereas they do not need to add all the species, reactions, and their relation. On the other hand, such an approach requires more manual effort and could possibly result in more inappropriate structures recommended, because of the few restrictions.

Therefore, the element names can be considered in addition, which constitutes the second approach. An one-to-one-comparison seems not very promising, because of the variety of possible options to name entities. Another option is to use a string edit distance metric or to compare the names by their sets of synonyms extracted from a dictionary. A string edit distance may fail, if words are syntactically very close, but semantically not even similar. Using synonyms is problematic for names that can be interpreted by humans, but are not included in a common dictionary. An example could be a sequence of abbreviations or chemical entities. However, both methods are used in several existing application cases and cope the problems of natural language ambiguity to some extent.

The third approach focuses on the semantic annotations. In contrast to names, the ontology concepts annotated on the network elements are unambiguous. Even though their importance was rated controversial in the survey (see Section 3.3), the common scheme of an ontology facilitates an improved automated processing and precise similarity measures. Integrating ontology terms in a recommender system may support the more widespread use of them.

The first method to compare two network elements by their annotations is to pairwise check the equivalence of all their resource URIs with the corresponding biological qualifiers. This method is quite strict, but rather simple to calculate. Furthermore, it is possible in this way to efficiently search for matching entities in MaSyMoS, because the URIs are stored as resource property and can directly be used for queries.

The other method is not considered to search for matches in MaSyMoS, but can measure the degree of similarity between ontology concepts annotated on network elements. There are several ontologies used in the domain of systems biology. Considering these separate ontologies, a comparison is only possible for ontology concepts within one ontology. To compare two concepts, the shortest path between them is calculated representing their distance in the ontology. In addition, the depth of both concepts in the ontology is considered, because deeper, more concrete terms are assumed to be more similar. Referencing to the evaluation results in Li et al. (2003), a non-linear combination of these two measures is assumed to constitute a precise similarity measure.

A problem of this approach is that possibly several concepts of different ontologies can be annotated on each network entity, but only concepts of the same ontology can be compared under the given conditions. Therefore, for each resource of an entity all resources of the other entity must be considered. Doing this for all entities in two networks is quite complex. Even annotations describing the same semantics can point to different resources. A further problem of the approach is that several concepts of the same

ontology can be annotated on a network entity. Even if the similarity for all concepts is calculated pairwise and one would, for example, consider the best matching concepts for further calculation, it is not clear how to assemble an appropriate score. To demonstrate this, the existence of three species s_1 , s_2 , and s_3 with several annotated resources is assumed, whereby s_1 and s_2 shall be compared to s_3 . If s_1 has one exact match with a resource of s_3 , but s_2 has two resources annotated that are quite similar to two resources of s_3 , which of the both species is the better match? Furthermore, considering different biological qualifiers would be necessary. One option to handle both problems regarding several annotated resources, is to restrict the considered ontology only to SBO, because each network entity can have only a single corresponding SBO-concept and there are no different biological qualifiers to consider.

Names and annotations are both optional in SBML, which is why they can be omitted. Furthermore, names can be meaningless or automatically generated, but may provide additional information, not provided by semantic annotations. Ontologies in contrast, provide common schemes that define semantics unambiguously. Therefore, a combination of the both previous approaches is promising. Either labels can be considered for similarity calculation, if an entity is not annotated, or a similarity measure based on a weighted combination of both approaches can be used.

Summarising, calculating the similarity of entities by means of the distance in the ontology is very complex. In this thesis, the one-to-one comparison of URIs is preferred for the strategies, because it can efficiently be utilised to query MaSyMoS. Other methods suggested could extend the strategies in the future.

5.3 Strategy 1: based on a single entity

The first strategy recommends for a selected network vertex another vertex with the corresponding edge between them. More precisely, recommendations for a reaction are the species, which are most likely to participate in the reaction, with an edge representing the role as reactant, modifier, or product. Recommendations for a selected species are in turn the reactions, in which the species may participate. The user can select a certain entity or the last one added is selected automatically. The calculation of recommendations is based on the resource URIs annotated on the selected entity together with the corresponding biological qualifier specifying the relation. Equivalent entities are searched in MaSyMoS. The result is a list of the immediate neighbours with the corresponding edge type. These neighbours are the species or reactions, which are connected to the given entity in any of the biological networks in MaSyMoS. They are characterised by one annotated resource URI with the corresponding biological qualifier and the occurrence frequency of the entity regarding the URI and qualifier. This means, if a neighbour in MaSyMoS has two resources annotated, it corresponds to two separate recommendation entries. This is, because several URIs can occur with different frequencies, and in this way the system provides the user the most common term. The results are ranked according to the frequency of joint occurrence or number of models in which they occur.

In the same way, it is possible to search for entities based on equivalent naming, instead of annotations. Furthermore, in future work similar entities could be searched by using

a threshold for a string edit distance between the names or for a distance between the ontology concepts.

5.4 Strategy 2: one-to-one match of the whole network

This strategy recommends single entities in the same way as in the previous strategy, but based on the whole network currently developed. The user can select a certain entity, for which neighbours shall be recommended. All structures equivalent to the network under construction are searched in MaSyMoS, assuming that the entity selected by the user is connected to a further entity. Like in strategy 1, the immediate neighbours of the selected entity are the recommendations, together with the corresponding edge type, one annotated resource URI, the biological qualifier and the occurrence frequency. A ranking is possible according to the frequency of joint occurrence or the number of models in which they occur.

To search for equivalent networks in MaSyMoS, two options are considered. The first option is to only consider the structure of the network under construction. The second option incorporates the annotations of the network elements. This may provide more precise recommendations than the first option, but bears the risk that no elements are retrieved for recommendation, because of the severe restrictions of an one-to-one-match. This strategy can also be adapted to only consider a certain type of ontology, for example SBO. Furthermore, it can be extended by calculating the edges for the recommended entities to all other vertices in the currently developed network.

5.5 Strategy 3: frequent patterns

The advantage of this strategy is that it facilitates the recommendation of larger network fragments. The idea is to extract frequent network patterns from the database, measure their similarity to the network under construction, and recommend similar patterns. The pattern extraction is performed by a frequent subgraph mining algorithm, such as gSpan (Yan and Han, 2002). This extracts all the subgraphs that occur in at least a given number of graphs. The applicability for networks stored in MaSyMoS is demonstrated by Lambusch (2015). One option for the pattern generation is to extract only structural patterns, which means graphs only with their types of vertices and edges. Another option is to use the ontology concepts. Because frequent subgraph mining algorithms use the NP-hard subgraph-isomorphism testing, an appropriate mediocrity must be found between computational complexity and restrictions of the semantics. If only structural patterns are mined, the computation is complex in particular, but if distinguishable semantic is incorporated, the resulting patterns may be quite individual cases with low frequency of occurrence. Therefore, the second option chosen is to use ontology concepts restricted to SBO.

All extracted patterns are candidates for the recommendation. The final fragments recommended are the patterns with minimum distance to the currently developed network. The distance metric is crucial for this strategy. It is assumed that the best results can be achieved by a combination of similarity measures for structural and semantic comparison. The structures can be compared by considering, for example, the total number

of nodes and edges or the number of common entities. For the semantic similarity, entities can be compared pairwise by methods described in Section 5.2 and an entire score can be computed by summing the score of the pairwise comparison. Furthermore, combined methods like graph matching as in Yang and Sze (2007) can be used. However, the aforementioned methods are quite complex.

Here, an one-to-one matching by means of MaSyMoS is considered further. The extracted patterns can be stored in this database and then the network under construction is used as a query. The recommendations then are the patterns containing the developed network, but having further nodes and edges. These entities can then be integrated in the network under construction. The recommendations can be ranked according to the frequency of the patterns in the database or the number of the extending entities. This strategy is also adaptable for names instead of SBO concepts. Then, string edit distance can be used to calculate the similarity.

5.6 Strategy 4: semantically similar models

This strategy recommends models, which contain a portion or all of the resource URIs extracted from the model under construction. For this purpose, all URIs present in the model are collected. By means of a Cypher query, all entities in MaSyMoS are searched that are annotated with at least one of the given URIs. The query outcome are the corresponding models with a list of the respective entities, their number, the total number of matching URIs in the model, and the proportion of matching entities to matching URIs. The models are used as recommendations and ranked according to the number of matching URIs. For the case, where several recommendation entries have the same number of matching URIs, the results are further ranked according to the aforementioned proportion. The strategy can be adapted to let the user select the URIs of the model currently developed, which shall be matched.

5.7 Strategy 5: common subgraph

This strategy combines ideas of strategies 3 and 4. Networks shall be retrieved, which contain a subgraph of the network under construction, called the common subgraph. By using subgraph isomorphism testing, it is computationally too intensive to calculate the maximum common subgraphs for all models in the database. Therefore, the semantically similar models are searched first according to strategy 4. The top five models are then used to find the maximum common subgraph regarding the network currently developed. This can be done by means of a subgraph mining algorithm as mentioned in strategy 3. For each of the top five models, the elements surrounding the common subgraph are recommended. An option to do this is to recommend for each entity of the common subgraphs the sequences of the next three neighbours. The sequences, which occur frequently, may connect several nodes of the common subgraph by means of intermediate extending neighbours. Thus, such sequences are ranked higher, assuming that users want to extend their networks with entities connecting several nodes already present.

5.8 Strategy 6: entity list and edge estimation

The idea of this strategy is to provide users with complex extensions, where relations to arbitrary entities of the model under construction are considered. To this end, a list of species and reactions is recommended, where the user can choose all entities that shall be integrated in the model. The second step computes the edges between all entities. To recommend a list of entities two options are considered.

The first option lets users choose entities of their model, for which neighbours shall be searched. For each selected entity, strategy 1 is performed, but instead of searching only for the direct neighbours, all neighbours reachable in a user-defined number of steps are retrieved from MaSyMoS. The resulting entity lists for all selected entities are merged and recommended to the user.

The second option is based on strategy 4, where semantically similar models are searched according to the number of entities that match URIs present in the currently developed model. For the matching entities of the top five models, all neighbours reachable in a user-defined number of steps are then retrieved from MaSyMoS. The resulting entity lists for the five models are merged and recommended to the user.

The users can choose all entities from the list that shall be integrated in their model. Afterwards all edges are estimated by a network completion or reconstruction algorithm (see Section 2.5).

6 Implementation

The recommendation strategies resulting from this thesis are envisioned to constitute an improvement within the modelling of biological networks. Therefore, an implementation should extend an available modelling tool. The survey results (see Section 3.3) and tool characteristics (see Section 2.4) indicate that an extension of CellDesigner seems reasonable. The first section gives more information about the possible use of CellDesigner to include the proposed recommendation strategies. In future work, the information used as a basis for the recommendation strategies shall be extracted from the network that a user creates in CellDesigner.

The developed strategies are all based on querying the model database MaSyMoS. The queries shall be created according to the information about the nodes, edges, and annotations extracted from the user's currently developed model. Beginning from Section 6.2, exemplary Cypher queries for several of the developed strategies are described, assuming that the required information about the user's model under construction are given. Appendix C shows the corresponding Cypher queries and the results for these exemplary queries are described in Chapter 7.

6.1 CellDesigner plug-in

CellDesigner facilitates the graphical modelling of networks and is SBML-compliant (Funahashi et al., 2003). The software version 4.0 and above can be extended by simply writing a Java program and adding it as plug-in⁷. Then, the plug-in can be called within

⁷<http://www.celldesigner.org/plugins.html>

CellDesigner from the menu. The CellDesigner Plugin API documentation⁸ provides information about the interfaces and classes that can be used for the development of a plug-in. Using the *CellDesignerPlugin*-class, it is possible to get the objects present in the user’s current model, for example, the species and reactions. This can be used to get the data needed to perform the database queries for the strategies proposed in this thesis. The class includes methods to get selected nodes, such as *getSelectedSpeciesNode()*, *getSelectedReactionNode()*, *getSelectedAllNode()*, *getAllSpeciesNodes()*, or *getAllReactionNodes()*. The classes *PluginSpecies* and *PluginReaction* provide the opportunity to retrieve the entity details required for the strategies of this thesis. The corresponding methods for both classes are *getId()*, *getName()*, and *getAnnotation()*. Similar to the structure of an SBML-file (see Appendix B.1), the network’s edges are only defined for reactions having a list of participating reactants, modifiers, and products. In CellDesigner, the *PluginReaction*-class provides methods, such as *getListOfReactants()* or *getNumReactants()*, which are similarly available for modifiers and products. These can be used to create a graph structure compatible with the one of MaSyMoS. Then, the Cypher queries could be created and executed within the source code of the developed CellDesigner plug-in. The resulting information retrieved from MaSyMoS could in turn be processed by the plug-in and shown to the user as recommendations. The integration of recommendations chosen by a user could be done in the CellDesigner plug-in by adding or modifying the user’s current network, for example, with the method *addReactant(PluginSpeciesReference ref)* of the *PluginReaction*-class. The following sections describe exemplary Cypher queries for several of the developed strategies (see Chapter 5), assuming that the required information about the user’s model under construction are already given.

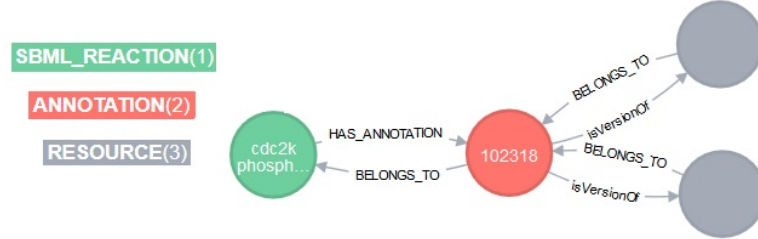
6.2 Strategy 1: based on a single entity

In this strategy, all direct neighbours for one selected entity with its annotated resources shall be found. It is assumed that the following information about the selected entity can be extracted from the user’s model in CellDesigner: the node type (reaction or species), all associated resource URIs, and the biological qualifier representing the entity’s relations to the URIs. Figure 6 shows an example for a reaction, which a user could have selected, and its annotated resources to query MaSyMoS according to strategy 1. Listing 6.2 shows the associated Cypher query to retrieve the data for recommendations. Appendix C contains the Cypher queries for all strategies that are prototypically implemented.

The Cypher query contains a *MATCH*-clause describing the structure of the given graph. This means, for the example, a reaction node having a connected annotation node that assigns two resources to the reaction, both by the biological qualifier *isVersionOf*. Furthermore, the *MATCH*-clause has to include the structures searched. In the example case, it is every species, which is directly connected to the given reaction and has annotated a resource. In addition, the model containing the matched reaction is searched. A *WHERE*-clause defines the given values for the properties, namely the both URIs corresponding to the both resources given for the reaction. One entry in

⁸<http://www.celldesigner.org/plugin/pluginAPI44/>

Figure 6: Exemplary input structure for strategy 1



The figure shows for one reaction, which a user could have selected, the corresponding structure that is relevant to query the database for strategy 1. More precisely, a reaction with two annotated resources is shown.

the query result shall correspond to one URI of a direct neighbour with a certain biological qualifier representing the relation. Therefore, the *RETURN*-clause of the query defines the result set by the edge type connecting the reaction to the found neighbour, further details about the neighbour, and the numbers of matching models and neighbours found. The further information about the neighbour include the node type, the qualifier connecting the neighbour to its resource, and the resource URI. By means of an *ORDER BY*-clause, the results are ranked according to the numbers of matching models and species. The query results created in this way already provide the data as it shall be recommended for strategy 1 (see Section 7.1). It is only necessary to receive them in the CellDesigner plug-in and present the table to the user appropriately.

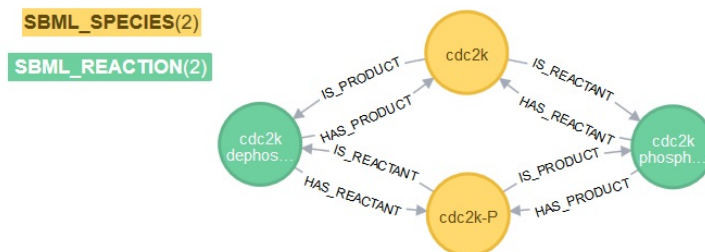
Listing 1: Exemplary Cypher query for strategy 1

```
MATCH
(reac:SBML_REACTION)-[:HAS_ANNOTATION]->(ar:ANNOTATION),
(ar)-[:isVersionOf]->(r1:RESOURCE),
(ar)-[:isVersionOf]->(r2:RESOURCE),
(reac)-[rel]->(s:SBML_SPECIES),
(s)-[:HAS_ANNOTATION]->(as:ANNOTATION)-[qualifier]->(rs:RESOURCE),
(m:SBML_MODEL)-[:HAS_REACTION]->(reac)
WHERE
r1.URI=~".*2.7.11.1" OR r2.URI=~".*G0:0006468"
RETURN
TYPE(rel) as edgeType, labels(s) as nodeType,
TYPE(qualifier) as qualifier, rs.URI as URI,
COUNT(DISTINCT ID(m)) as numberOfModels,
COUNT(DISTINCT s) as frequency
ORDER BY numberOfModels DESC, frequency DESC
```

6.3 Strategy 2.1: one-to-one match of the network structure

The first option of strategy 2 searches for direct neighbours of one selected entity based on the whole network, but only considers the structure of the user's currently developed network without annotations. Appendix C.2 shows an exemplary Cypher query for this strategy. It is assumed that the whole network structure can be extracted from the user's model in CellDesigner. An example for a network under construction is shown in

Figure 7: Exemplary input structure for strategy 2.1



The figure shows the structure of an exemplary network currently developed as it is relevant to query the database for strategy 2.1. More precisely, two reactions are shown that are connected in a cyclic way by having the product of each other as reactant.

Figure 7. The given and the searched structure are again defined in the *MATCH*-clause. The given structure is the whole network shown in the figure. It is assumed that the user has selected one of the both species for recommendations. Then, the searched structure is this species directly connected to a further reaction, which has an associated resource. The model containing the reaction shall also be searched. No *WHERE*-clause is needed, because there are no further restrictions for the structure. The *RETURN*-clause is the same as for strategy 1 and the results can also directly be used as recommendations.

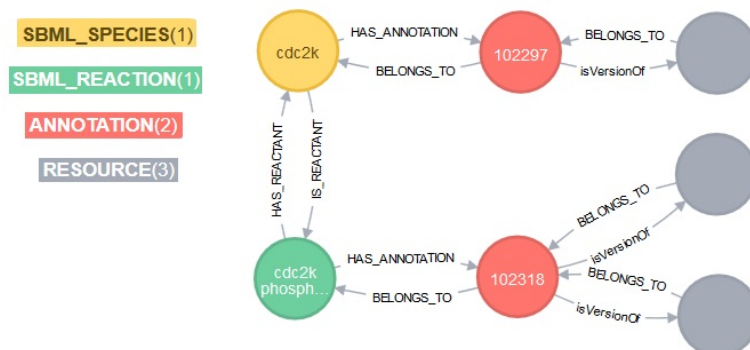
6.4 Strategy 2.2: one-to-one match of the network with annotations

Strategy 2 option 2 finds direct neighbours for one selected entity based on the whole network and considers the structure together with annotations. Appendix C.3 shows an example for a Cypher query to perform this strategy. It is assumed that the network structure and the entities' associated resources can be extracted from the user's model in CellDesigner. An exemplary network under construction is shown in Figure 8. In the example, the given structures are the connected reaction and species with their associated resources. It is assumed that the user has selected the species to extend it. Thus, the exemplary query contains a relation from the species to a further reaction. The query is similar to the one for strategy 2.1, but defines the URIs of the given structures in the *WHERE*-clause.

6.5 Strategy 3: frequent patterns

For this strategy, frequent network patterns must be extracted from the set of available SBML-models first, which can then be used as recommendations. In Lambusch (2015) a frequent subgraph mining algorithm is used to search for structural patterns by means of models stored in MaSyMoS. Thereby, the types of vertices and edges are regarded for species and reactions, but the associated annotations are not considered. For the curated SBML-models of BioModels Database release 26, 53 frequent patterns are found, which occur in at least 250 of 453 input graphs. These structural patterns can be used as recommendations. Another option is to proceed as in Lambusch (2015), but incorporate

Figure 8: Exemplary input structure for strategy 2.2



The figure shows the structure of an exemplary network currently developed as it is relevant to query the database for strategy 2.2.

More precisely, a connected species and reaction are shown. The species has one annotated resource and the reaction has two, respectively.

the SBO-terms in the networks. This reduces the number of models that can be used for the frequent subgraph mining algorithm by around 80 percent. Because of restrictions for the input format used to perform the frequent subgraph mining algorithm, it is not possible to incorporate all resources associated with the available networks.

The frequent patterns can be stored in MaSyMoS to use them for recommendations. Then, it is possible to search for the patterns in the database by means of Cypher queries similar to those shown in Appendix C. To make such frequent patterns available in MaSyMoS is a task for future work, because an approach to retrieve larger, more significant patterns has to be found first. All patterns that are found by a one-to-one matching with the user's network under construction and have at least one more entity can be recommended. If a user chooses a recommended pattern, the additional entities can be integrated in the user's model. The recommended patterns can be ranked according to the frequency of their occurrence or the number of the extending entities.

6.6 Strategy 4: semantically similar models

This strategy searches for models, which contain a portion or all of the resource URIs present in the model under construction. The network structure and the qualifiers for the resources are not considered. All URIs present in the model are collected and all entities in MaSyMoS are searched that are annotated with at least one of the given URIs. Appendix C.4 shows an exemplary Cypher query for this strategy, assuming that the network currently developed complies with the one shown in Figure 8. The *MATCH*-clause of the query only consists of a model containing an entity, which is annotated with a resource. In the *WHERE*-clause the restrictions for the resource URI are defined. The example network contains three resources. Thus, models are searched that have an entity annotated with at least one of the three resource URIs. The query outcome (*RETURN*-clause) are the corresponding models with a list of the respective entities, their number, the total number of matching URIs in the model, and the proportion of

matching entities to matching URIs. For the models, the internal database ID and the name are retrieved, because only the internal database ID is unique among all stored models and prevents the merge of several result entries. The models are ranked according to the number of matching URIs and further to the aforementioned proportion. The query results provide the data as it shall be recommended for strategy 4 (see Section 5.6) and only need to be presented in the CellDesigner plug-in appropriately.

6.7 Strategy 5: common subgraph

With this strategy networks shall be found, which contain a subgraph of the network under construction, called the common subgraph. First, the semantically similar models are searched according to strategy 4. The top five models are then used to find the maximum common subgraph regarding the network currently developed. This can be done by means of the tool used in Lambusch (2015). Then, the common subgraph can be found regarding only the structure without annotations. All elements surrounding the common subgraph can be recommended. As mentioned for strategy 3, the input format for the tool used to find common subgraphs is restricted, which makes it barely feasible to incorporate comprehensive information about the networks' semantics. The implementation of this strategy is a task for future work, because possibilities to use a less restricted input format for finding the common subgraph have to be examined first.

6.8 Strategy 6: entity list and edge estimation

The idea of this strategy is to provide users a list of entities, where they can choose all reactions and species that shall be integrated in the model. Afterwards, the edges are estimated. For the first option, a user has to choose entities of the model under construction, for which neighbours shall be searched. Then, strategy 1 is performed for each selected entity, but with an extended query that searches for all neighbours reachable in a user-defined number of steps. The resulting entity lists for all selected entities are merged and recommended to the user. This option combines parts of the queries for strategy 1 and the following strategy 6.2. and is not considered further.

The second option uses strategy 4 to search the top five of semantically similar models according to the entities, which have at least one of the resources annotated that are contained in the network under construction. For these entities, all neighbours reachable in a user-defined number of steps are retrieved and recommended to the user as an entity list. The network currently developed is for this example again the one shown in Figure 8 and it is assumed that strategy 4 is already performed with the query shown in Appendix C.4. The top five models resulting from the execution of this query (see Section 7.4) are used for the strategy described here. The corresponding query is shown in Appendix C.5. The first *MATCH*- and *WHERE*-clause are similar to the one of strategy 4, but restrict the internal model IDs to those of the top five models. A second *MATCH*-clause searches for neighbours reachable in three steps from one of the matching entities, which are found by means of the first part of the query. For these neighbours also their resources are searched. The result is a list of entities, which are each characterised by their type, the biological qualifier connecting it to a resource, the

corresponding resource URI, the frequency with which they occur as a neighbour, and the number of models in which they occur. The results are ranked by the frequency of occurrence and the number of models.

The query results provide the data as it shall be recommended. The CellDesigner plugin should present the entities in a way, where the users can tick off all those that shall be integrated in their model. The second step of this strategy shall estimate the edges between all entities by means of a network completion or reconstruction algorithm. This step is not yet implemented, because available tools (see Section 2.5 for an example) require specific input types. Examining the applicability of existing tools for the available data is a task for future work.

7 Results

All developed strategies are based on querying the model database MaSyMoS. Appendix C shows for several of the developed strategies exemplary Cypher queries, for which the results retrieved from the database are examined in this chapter. For all examples, it is assumed that the user wants to model the network of *Tyson1991 - Cell Cycle 6 var* shown in Figure 1. This builds a reference point of what the user might expect as results. Chapter 6 describes the assumed initial situation for every query, which means the data used by the respective strategy.

7.1 Strategy 1: based on a single entity

This strategy finds all direct neighbours for one selected entity in the user's current network. Appendix C.1 shows the exemplary Cypher query, for which the description can be found in Section 6.2 and the results are shown here. It is assumed that the network under construction is the one shown in Figure 7 with all corresponding annotations from the original model *Tyson1991 - Cell Cycle 6 var*. The user has selected the reaction named *cdc2k phosphorylation*, for which Figure 6 shows the structure used in the query. The reaction has two resources, both are annotated by the qualifier *isVersionOf*. One links to the term *Non-specific serine/threonine protein kinase*, the other to the term *protein phosphorylation*. The table resulting from the execution of the query is shown in Figure 9. The figure shows the top ten results for the query. In total, 286 results were found, which means 286 existing combinations of an edge type, a qualifier, and an URI for a possible neighbour. The query took less than a second to complete. The maximum number of models, in which a possible neighbour for the given reaction is contained, is seven. The maximum frequency, with which a certain neighbour occurs, is 25. The top ten results contain five URIs, assigned to a species by different qualifiers or the species is connected to the given reaction by different edge types. These five URIs represent the terms *Mitogen-activated protein kinase 1*, *Phosphoprotein*, *Dual specificity mitogen-activated protein kinase kinase 1*, *Period circadian protein*, and *polypeptide chain*. Neighbours that are contained in *Tyson1991 - Cell Cycle 6 var* can only be found in the result set at position 24 and above with a maximum of two models containing them. Entries at position 102 and above are only contained in one model and also with a frequency of one.

Figure 9: Exemplary query result for strategy 1

edgeType	nodeType	qualifier	URI	numberOfModels	frequency
HAS_PRODUCT	[SBML_SPECIES]	isVersionOf	http://identifiers.org/uniprot/P26696	7	12
HAS_PRODUCT	[SBML_SPECIES]	isVersionOf	http://identifiers.org/kegg.compound/C00562	5	8
HAS_REACTANT	[SBML_SPECIES]	hasPart	http://identifiers.org/uniprot/Q05116	4	25
HAS_REACTANT	[SBML_SPECIES]	hasPart	http://identifiers.org/uniprot/P26696	4	19
HAS_PRODUCT	[SBML_SPECIES]	isVersionOf	http://identifiers.org/uniprot/P07663	4	10
HAS_REACTANT	[SBML_SPECIES]	isVersionOf	http://identifiers.org/uniprot/P07663	4	10
HAS_PRODUCT	[SBML_SPECIES]	isVersionOf	http://identifiers.org/uniprot/Q05116	4	7
HAS_PRODUCT	[SBML_SPECIES]	HAS_SBOTERM	urn:miriam:biomodels.sbo:SBO:0000252	4	5
HAS_REACTANT	[SBML_SPECIES]	HAS_SBOTERM	urn:miriam:biomodels.sbo:SBO:0000252	4	5
HAS_MODIFIER	[SBML_SPECIES]	HAS_SBOTERM	urn:miriam:biomodels.sbo:SBO:0000252	3	7

Returned 286 rows in 769 ms.

The figure shows the first ten results for the exemplary Cypher query in Appendix C.1. The results are retrieved from MaSyMoS. They represent exemplary recommendations for strategy 1.

7.2 Strategy 2.1: one-to-one match of the network structure

The first option of strategy 2 finds all direct neighbours for one selected entity based on the user's whole network, regarding only the structure without annotations. Appendix C.2 shows the exemplary Cypher query corresponding to this strategy. Section 6.3 describes the query, for which the results are shown here. The example for the network under construction is again the one shown in Figure 7 with all corresponding annotations from the original model *Tyson1991 - Cell Cycle 6 var*, but this strategy uses only the structure for the search. The user has selected one of the both species for extension. Because of the cyclic structure, which is equal for both species, it is not relevant for the query in strategy 2.1, which one is chosen. The resulting table is shown in Figure 10. The figure shows the top ten results for the query. In total, 3186 existing combinations for a possible neighbour of a species contained in a cyclic structure were found. The query took nearly 43 seconds to complete, which is noticeably long. The maximum number of models containing a possible neighbour is 31 and the maximum frequency of occurrence is 360. The top ten results contain mainly reactions, for which the given species takes a role as product. The found qualifiers for the top ten results are only *HAS_SBOTERM* and *isVersionOf*. The top ranked term represents *phosphorylation*. Furthermore, most of the other terms in the top ten results link to similar terms, related to (de)phosphorylation, (dis)assembly, or dissociation, sometimes more general and sometimes referring to a protein or protein complex. Result entries at position 1921 and above are only contained in one model and also with a frequency of one.

7.3 Strategy 2.2: one-to-one match of the network with annotations

The second option of strategy 2 finds all direct neighbours for one selected entity based on the user's whole network incorporating the annotations. Appendix C.3 shows the

Figure 10: Exemplary query result for strategy 2.1

edgeType	nodeType	qualifier	URI	numberOfModels	frequency
IS_PRODUCT	[SBML_REACTION]	HAS_SBOTERM	urn:miriam:biomodels.sbo:SBO:0000216	31	93
IS_PRODUCT	[SBML_REACTION]	isVersionOf	http://identifiers.org/go/GO:0006468	29	80
IS_PRODUCT	[SBML_REACTION]	isVersionOf	http://identifiers.org/go/GO:0043241	28	206
IS_REACTANT	[SBML_REACTION]	isVersionOf	http://identifiers.org/go/GO:0006461	28	125
IS_PRODUCT	[SBML_REACTION]	isVersionOf	http://identifiers.org/go/GO:0006461	27	156
IS_PRODUCT	[SBML_REACTION]	isVersionOf	http://identifiers.org/go/GO:0006470	27	83
IS_PRODUCT	[SBML_REACTION]	HAS_SBOTERM	urn:miriam:biomodels.sbo:SBO:0000180	25	146
IS_PRODUCT	[SBML_REACTION]	HAS_SBOTERM	urn:miriam:biomodels.sbo:SBO:0000330	25	63
IS_REACTANT	[SBML_REACTION]	HAS_SBOTERM	urn:miriam:biomodels.sbo:SBO:0000179	24	89
IS_PRODUCT	[SBML_REACTION]	isVersionOf	http://identifiers.org/go/GO:0006412	24	47

Returned 3186 rows in 42519 ms, displaying first 1000 rows.

The figure shows the first ten results for the exemplary Cypher query in Appendix C.2. The results are retrieved from MaSyMoS. They represent exemplary recommendations for strategy 2.1.

exemplary Cypher query, which is described in Section 6.4. In this section, the results of the query are examined, which represent possible recommendations for strategy 2.2. The considered example for the network under construction, which also represents the structure and resources as they are used for the query, is the one shown in Figure 8. The user has selected the species for extension. The table resulting from the query is shown in Figure 11. The figure shows all results for the query, only four possible neighbours for the species were found. The query took a half second to complete. Each possible neighbour occurs only once for the given restrictions. All four entries represent a reaction, for which the given species takes a role as product. The retrieved qualifiers are *isVersionOf* and *hasVersion*. All found neighbours stem only from the network of *Tyson1991 - Cell Cycle 6 var*, corresponding to the reactions *cdc2k dephosphorylation*

Figure 11: Exemplary query result for strategy 2.2

edgeType	nodeType	qualifier	URI	numberOfModels	frequency
IS_PRODUCT	[SBML_REACTION]	isVersionOf	http://identifiers.org/obo.go/GO:0006470	1	1
IS_PRODUCT	[SBML_REACTION]	isVersionOf	http://identifiers.org/ec-code/3.1.3.16	1	1
IS_PRODUCT	[SBML_REACTION]	isVersionOf	http://identifiers.org/obo.go/GO:0000079	1	1
IS_PRODUCT	[SBML_REACTION]	hasVersion	http://identifiers.org/reactome/REACT_6308	1	1

Returned 4 rows in 544 ms.

The figure shows all four results obtained for the exemplary Cypher query in Appendix C.2. The results are retrieved from MaSyMoS. They represent exemplary recommendations for strategy 2.2.

or *cyclin_cdc2k dissociation* shown in Figure 3. The one-to-one-match of this strategy is very strict. Even for a structure with only two entities and three resources in total, just a few results can be retrieved.

7.4 Strategy 4: semantically similar models

This Strategy finds all models, which contain at least one of the resource URIs present in the user's currently developed network. Appendix C.4 shows an exemplary Cypher query for this strategy. Section 6.6 describes this query and this section its results retrieved from MaSyMoS. The result set can be used as a list of recommendations for the considered example. It is assumed that the user's network with its annotated resources is the one shown in Figure 8. The three associated resource URIs are the basis for recommendations in strategy 4. For the reaction, one URI links to the term *protein phosphorylation*, the other to the term *Non-specific serine/threonine protein kinase*. The species has one associated resource URI linking to the term *Cyclin-dependent kinase 1*. The query results are shown in Figure 12.

Figure 12: Exemplary query result for strategy 4

internalID	modelName	matchingEntityNames	matchingURIs	numEntities	numURIs	proportion
102286	Tyson1991 - Cell Cycle 6 var	[deactivation of cdc2 kinase, cdc2k phosphorylation, cdc2k phosphorylation, total_cdc2, cdc2k-P, cdc2k, p-cyclin_cdc2-p, p-cyclin_cdc2]	[http://identifiers.org/obo.go/GO:0006468, http://identifiers.org/ec-code/2.7.11.1, http://identifiers.org/uniprot/P04551]	7	3	0.42857142...
105835	Kofahl2004_PheromonePathway	[(empty), (empty)]	[http://identifiers.org/ec-code/2.7.11.1, http://identifiers.org/go/GO:0006468]	2	2	1
104196	Goldbeter1995_CircClock	[first phosphorylation of PER, first phosphorylation of PER, second phosphorylation of PER, second phosphorylation of PER]	[http://identifiers.org/go/GO:0006468, http://identifiers.org/ec-code/2.7.11.1]	2	2	1
122372	Leloup1998_CircClock_LD	[PER-p phosphorylation, PER-p phosphorylation, TIM phosphorylation, TIM phosphorylation, TIM-p phosphorylation, TIM-p phosphorylation, PER phosphorylation, PER phosphorylation]	[http://identifiers.org/ec-code/2.7.11.1, http://identifiers.org/go/GO:0006468]	4	2	0.5
105090	Leloup1999_CircClock	[First Phosphorylation of TIM, First Phosphorylation of TIM, Second Phosphorylation of PER, Second Phosphorylation of PER, Second Phosphorylation of TIM, Second Phosphorylation of TIM, First Phosphorylation of PER, First Phosphorylation of PER]	[http://identifiers.org/go/GO:0006468, http://identifiers.org/ec-code/2.7.11.1]	4	2	0.5

Returned 62 rows in 1828 ms.

The figure shows the first five results for the exemplary Cypher query in Appendix C.4. The results are retrieved from MaSyMoS. They represent exemplary recommendations for strategy 4.

The figure shows the top five results for the query. In total, 62 models were found that contain entities with at least one of the mentioned URIs associated. The query took nearly 2 seconds to complete. Only the network of *Tyson1991 - Cell Cycle 6 var* contains all three URIs and seven of its entities are associated with at least one

Figure 13: Exemplary query result for strategy 6.2

nodeType	qualifier	URI	freqRelated	numOfModels
[SBML_SPECIES]	isVersionOf	http://identifiers.org/uniprot/P07663	91	3
[SBML_REACTION]	isVersionOf	http://identifiers.org/ec-code/3.1.3.16	63	4
[SBML_SPECIES]	isVersionOf	http://identifiers.org/uniprot/P49021	63	2
[SBML_REACTION]	isVersionOf	http://identifiers.org/ec-code/2.7.11.1	46	5
[SBML_SPECIES]	isVersionOf	http://identifiers.org/kegg.compound/C00562	38	1
[SBML_REACTION]	hasVersion	http://identifiers.org/reactome/REACT_6308	34	1
[SBML_REACTION]	isVersionOf	http://identifiers.org/obo.go/GO:0006470	33	1
[SBML_REACTION]	isVersionOf	http://identifiers.org/go/GO:0006468	32	4
[SBML_REACTION]	isVersionOf	http://identifiers.org/go/GO:0006470	30	3
[SBML_REACTION]	isVersionOf	http://identifiers.org/obo.go/GO:0006468	30	1

Returned 67 rows in 2144 ms.

The figure shows the first ten results for the exemplary Cypher query in Appendix C.5. The results are retrieved from MaSyMoS.

They represent exemplary recommendations for strategy 6, but for the integration of the entities an estimation of suitable edges is necessary.

of them. Several models contain two of the URIs, but 47 models contain only one. The maximum number of matching entities is 20 for a model that has two matching URIs. The corresponding model is not contained in the top five results, because its proportion of matching entities to matching URIs is too small. Using the proportion as second ranking criterion prevents models with a large amount of matching entities to be ranked higher than models, for which less entities match the same number of URIs.

7.5 Strategy 6.2: entity list based on strategy 4 and edge estimation

The idea of strategy 6 is to recommend a list of entities, where users can choose all reactions and species that shall be integrated in the model. Afterwards, the edges for the network are estimated. Option 2 of this strategy uses the top five of similar models retrieved from executing strategy 4, whereby also the entities are found that are associated with the given URIs. For these entities, all neighbours reachable in a user-defined number of steps are retrieved and recommended to the user. The network currently developed is for this example again the one shown in Figure 8. Appendix C.5 shows an exemplary Cypher query for this strategy, which is explained in Section 6.8. It is based on the top five models shown in Figure 12. Here, the results are considered, which are retrieved from MaSyMoS and can be used as recommendations for the considered example. They are shown in Figure 13. The figure shows the top ten results for the query. In total, 67 results were found, which means 67 entities reachable in three steps from one of the matching entities of strategy 4. The query took around

two seconds to complete. The maximum number of models, in which a resulting entity is contained, is five. The maximum frequency, with which an entity is reachable in three steps, is 91. The top ten results contain three species and seven reactions. All but one of the ten entities are connected to their resource by the biological qualifier *isVersionOf*. The other entity has the qualifier *hasVersion*. The only entity contained in five models has the associated URI representing *Non-specific serine/threonine protein kinase*, which is already an URI used as input for the query. Such redundancies should be avoided in a future implementation. The results also contain the four entities obtained for the example in Section 7.3.

The integration of the recommended entities in the user’s model is not yet implemented. Examining the applicability of existing tools for the network estimation is a task for future work. Therefore, the results described here can be used as recommendations in terms of an entity list, where a user can choose species and reactions for integration, but in contrast to the results of the prior sections, they do not complete the recommendation process.

8 Discussion

The developed strategies demonstrate that the realisation of recommendations for the modelling of biological networks is feasible. The developed methods support users to perform reusing tasks based on their biological network under construction. The methods reduce the search space in the increasing number of available networks and guide users to the data that may be of interest to extend their model. Several recommendation strategies are presented that can fulfil different needs according to the modelling intention. Exemplary recommendations are shown in Chapter 7. For several strategies, recommendations selected by users could directly be integrated in their currently developed model. The results indicate the applicability and suitability of the proposed strategies. An evaluation from the perspective of the systems biology community is still required for a comprehensive assessment.

8.1 Findings

The proposed strategies refer to content-based recommender systems and transfer recommendation ideas from the domain of business process modelling to the biological network modelling. To regard domain knowledge and provide more reliable recommendations, different approaches are considered to incorporate the semantics of biological networks (see Section 5.2). Each approach has its advantages and drawbacks. The names of the network entities, for example, are not restricted in SBML and users are free to incorporate additional context information, whereas the possible ambiguity makes an automated processing with high accuracy very difficult. For the exemplary implementation of the proposed strategies, the method of choice is an one-to-one matching of the network’s semantic annotations. This is quite strict, but rather simple and precise.

Strategy 1: based on a single entity

The exemplary recommendations for strategy 1 comprise 286 possible neighbours for the reaction named *cdc2k phosphorylation* based on its two associated resources. The query’s execution time of less than a second are gratifying. The top ten results (see Figure 9) contain species representing proteins, which is adequate according to the given reaction. Nevertheless, entities stemming from *Tyson1991 - Cell Cycle 6 var* occur at position 24 and above in the result list. This indicates that the context of the reaction selected by the user has to be incorporated, if quite specific entities are searched. Strategy 2 uses this context by an one-to-one matching of the whole network structure to recommend neighbours for an entity selected by the user.

Strategy 2: one-to-one match of the whole network

The first option for this strategy is to use information about the network structure without annotations. The exemplary results in Figure 10 show the first ten of 3186 possible neighbours for the species named *cdc2k* or its phosphorylated variant *cdc2k-P*. The large amount of results and the query’s execution time of nearly 43 seconds are not satisfactory. In consequence, this method should mainly be used, if the structure of the user’s network is more complex or significant. The problem may especially occur for cyclic structures, because rotation of the cycle results in the same structure. Thus, more neighbours are searched, such as in the example the once connected to *cdc2k* as well as neighbours for *cdc2k-P*. This can also cause higher frequency values.

The top ranked terms for the example include (de)phosphorylation, which exactly describes the two reactions contained in the assumed user’s network. Other terms are related to similar processes, such as (dis)assembly or dissociation. The terms are more general or protein-related. Although the top ranked URIs are not the same as used in *Tyson1991 - Cell Cycle 6 var*, the results demonstrate that the method is able to recommend semantically similar entities even by only considering the network structure. This may be only true, if the structure is representative. A problem is that the recommendations contain terms, which may describe the reactions in the cyclic structure, instead of only recommending further neighbours. This can be corrected in future implementations by adding restrictions to the queries describing the inequality of the given and searched neighbours. To reduce the search space, speed up the computations, and restrict the semantic for more accurate results, the strategy can also incorporate the networks’ annotations. The second option for strategy 2 is to match the network structure together with associated resource URIs one-to-one. All four results for the exemplary query are shown in Figure 11. All found neighbours stem only from the network of *Tyson1991 - Cell Cycle 6 var*. The few results for a network with only two entities and three associated resources indicate that this strategy is only helpful, if the user’s model under construction is very small or general, or the user wants to model exactly an already existing part of a model.

Because the first option imposes only a few restrictions and the second option is strongly restricted, a combination could be promising. For this purpose, the whole network structure could exactly be matched, while annotations are only considered for the entity selected by the user for extension. This can bring together the advantages of strategy

1 and 2 in future implementations.

Strategy 3: frequent patterns

The reason to use a frequent pattern approach, such as proposed for this strategy, is to support complex network fragments as recommendations. It is possible to use the structural patterns from Lambusch (2015) or to incorporate SBO-terms for the pattern extraction. Storing such patterns in MaSyMoS and query them for recommendations is feasible. The problem is to extract patterns, because related algorithms require subgraph isomorphism testing, which is a NP-hard problem. The results in the aforementioned thesis show that mainly small patterns, which are not actually diverse, can be found for the considered biological networks. Therefore, recommending such patterns may be not adequate. The authors of Li et al. (2014) use the same subgraph mining algorithm to recommend patterns for business process models, but remodel their process graphs to uniform models. In this way, they facilitate the extraction of large patterns. By remodelling, they can even find for the whole set of models the patterns that are only contained in a single network. Remodelling all biological networks considered in this thesis is barely feasible, especially regarding the increasing amount of available biological models. Furthermore, new models in the database require the maintenance of the pattern set, but algorithms for pattern extraction are not scalable. Because MaSyMoS facilitates fast subgraph isomorphism testing as used for the queries in Appendix C, it could be promising to integrate a pattern mining algorithm that can directly work with the database.

Strategy 4: semantically similar models

Strategy 4 finds all models, which contain at least one of the resource URIs present in the user's currently developed network. Figure 12 shows the first five of 62 recommended models for the exemplary query. The number of results and the execution time of nearly two seconds are satisfactory for three given URIs. For networks with considerably more associated resources both values may be much higher, because all models containing at least one of the given resource URIs are searched. Therefore, a future implementation should allow the users to state mandatory URIs and restrict the number of recommended models in this way. The strategy results fulfil the expectations by recommending the network of *Tyson1991 - Cell Cycle 6 var* first. This is, because it is the only model containing all three resource URIs. The second ranking criterion proposed for this strategy is the proportion of matching entities to matching URIs. As one can see in the figure, *Tyson1991 - Cell Cycle 6 var* has seven matching entities and accordingly the proportion value is smaller than the once of models with only two matching entities. The idea here is to recommend users primarily models with a high number of matching URIs, which are concentrated in a few entities. It has to be examined in future work, whether using this proportion is helpful for ranking.

Strategy 6: entity list and edge estimation

In this strategy, a list of entities is recommended and the user selects a set of those that shall be integrated in the model under construction. An example for recommended entities is shown in Figure 13, which shows the top ten of 67 results. These are the entities reachable in three steps from one of the matching entities of the top five models retrieved from performing strategy 4. The number of recommendations is manageable and the execution time of the query with around two seconds is satisfactory. The only entity contained in five models has an associated URI already used as input for the query. In future implementations this shall be avoided by stating inequality of the given and searched URIs. The top results also contain the four entities obtained for the example of strategy 2.2., which stem from *Tyson1991 - Cell Cycle 6 var*. This indicates that the ranking is adequate. The integration of the recommended entities in the user's model is not yet implemented. It remains an open question, if the integration of entities by estimation of the edges works accurately for the given situation and data. A problem might be that related algorithms, such as the one described in Section 2.5, estimate the structure based on the assumption that the network will then be complete. The applicability of these algorithms for networks, which are still under construction, has to be examined first.

8.2 Conclusion

The developed strategies demonstrate that methods of recommender systems are applicable to support users in modelling biological networks. Recommendations of existing model parts may ease the reuse by reducing the search space. The proposed strategies bring together search, comparison and integration of subnetworks. It remains open, whether recommendations facilitate a faster modelling process and what the consequences for the model quality are.

The proposed strategies are related to content-based recommender systems and utilise the user's currently developed model as the basis for recommendations. A main advantage is the user independence and transparency of the used methods. This means, no data about users is needed to recommend proper structures and it is possible to list features of the subnetworks that caused their recommendation.

The requirements defined in this thesis constitute a solid foundation, on which appropriate strategies could be developed. The survey results indicate that a model repository is an appropriate model source for reusing tasks and integrating recommendations in the tool CellDesigner is desirable. The proposed strategies use networks extracted from the 590 curated SBML-models available in the repository BioModels. Furthermore, they can extend CellDesigner by means of a plug-in. By extending CellDesigner, the compatibility to common modelling languages for biochemical reaction network is ensured. Further recommendation dependent and independent extension is possible by implementing additional plug-ins. The use of the database MaSyMoS as basis for recommendations facilitates the evolution of the recommendation source.

The exemplary implemented strategies seem to provide reasonable recommendations and the ranking can be further refined, if more information about the users' needs is available. It is assumed that no strategy can be preferred, but each may fulfil different

needs for different modelling intentions. Thus, the strategies should be implemented as one plug-in to let the users decide for a strategy appropriate for their current modelling situation. It is shown that the proposed methods are scalable by making use of the database MaSyMoS. Only strategy 2.1, which utilises just the network's structure, is not scalable. Nevertheless, this strategy may provide appropriate recommendations, if the network's structure is significant or rare. As the developed strategies demonstrate, it is feasible to recommend single network entities or semantically similar models adequately. In contrast, creating recommendations of complex fragments is quite difficult. This is mainly due to the complexity of the required algorithms, such as subgraph isomorphism or network completion.

Summary

Approaches to obtain recommendations of subnetworks during modelling were so far unexploited in systems biology. This thesis forms a solid basis for the development of a biological recommender system, which provides modellers diverse strategies to extend their networks. Important requirements are defined in terms of recommender systems in general and in particular for the application case of modelling biochemical reaction networks. The conducted survey revealed a first impression of how the modellers can be supported efficiently. The developed concept for generating recommendations comprises different semantic and structural approaches, which allow for utilising various features of the considered data basis. Several strategies are implemented and tested prototypically.

8.3 Future work

The next logical step is the comprehensive implementation of the proposed strategies to extend an existing modelling tool. A following evaluation by real users can reveal new insights in further requirements and the usability of the recommendation methods. An important topic for further research is the development of strategies to recommend complex subnetworks. The algorithms must be accurate, but also scalable. The integration of an algorithm for subgraph mining in MaSyMoS could reveal larger patterns that can be used as recommendations. There are several semantic approaches described in the concept, which were not considered for implementation. A task for future work is to develop further strategies based on the other semantic approaches, such as using the names contained in networks. Examining possible combinations of name similarity and the use of semantic annotations could reveal more comprehensive strategies.

Recommendations could be refined by letting users state their satisfaction or dissatisfaction. Furthermore, examining methods to incorporate information about users' behaviour, such as the number of times a recommendation is selected, could improve the quality of the recommender system. Then, a future direction could be to examine and integrate learning algorithms that use the aforementioned information.

A further analysis on how to improve the model quality by means of recommendations may be important. Possibly, an integration with tools for verifying syntactic correctness of models could be performed. The score for correctness should then influence the calculation for recommendations.

Bibliography

- Adomavicius, G. and Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering*, 17(6):734–749.
- Alm, R., Waltemath, D., Wolfien, M., Wolkenhauer, O., and Henkel, R. (2015). Annotation-based feature extraction from sets of sbml models. *Journal of biomedical semantics*, 6(1):1.
- Antoniotti, M., Policriti, A., Ugel, N., and Mishra, B. (2003). Model building and model checking for biochemical processes. *Cell biochemistry and biophysics*, 38(3):271–286.
- Blanchard, E., Harzallah, M., Briand, H., and Kuntz, P. (2005). A typology of ontology-based semantic measures. In *Proceedings of the Open Interop Workshop on Enterprise Modelling and Ontologies for Interoperability*, pages 13–14.
- Bobek, S., Baran, M., Kluza, K., and Nalepa, G. J. (2013). Application of bayesian networks to recommendations in business process modeling. In *Proceedings of the AI Meets Business Processes Workshop: 13th Conference of the Italian Association for Artificial Intelligence*, pages 41–50.
- Bunke, H., Foggia, P., Guidobaldi, C., Sansone, C., and Vento, M. (2002). A comparison of algorithms for maximum common subgraph on randomly connected graphs. In *Proceedings of the Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pages 123–132. Springer.
- Bunke, H. and Shearer, K. (1998). A graph distance metric based on the maximal common subgraph. *Pattern recognition letters*, 19(3):255–259.
- Butcher, E. C., Berg, E. L., and Kunkel, E. J. (2004). Systems biology in drug discovery. *Nature biotechnology*, 22(10):1253–1259.
- Courtot, M., Juty, N., Knüpfer, C., Waltemath, D., Zhukova, A., Dräger, A., Dumontier, M., Finney, A., Golebiewski, M., Hastings, J., et al. (2011). Controlled vocabularies and semantics in systems biology. *Molecular systems biology*, 7(1):543.
- Czauderna, T., Klukas, C., and Schreiber, F. (2010). Editing, validating and translating of sbgn maps. *Bioinformatics*, 26(18):2340–2341.

- Dijkman, R., Dumas, M., Van Dongen, B., Käärik, R., and Mendling, J. (2011). Similarity of business process models: Metrics and evaluation. *Information Systems*, 36(2):498–516.
- Fellmann, M., Zarvic, N., Metzger, D., and Koschmider, A. (2015). Requirements catalog for business process modeling recommender systems. In *Wirtschaftsinformatik Proceedings*, pages 393–407.
- Finkelstein, A., Hetherington, J., Li, L., Margoninski, O., Saffrey, P., Seymour, R., and Warner, A. (2004). Computational challenges of systems biology. *Computer*, 37(5):26–33.
- Funahashi, A., Morohashi, M., Kitano, H., and Tanimura, N. (2003). Celldesigner: a process diagram editor for gene-regulatory and biochemical networks. *Biosilico*, 1(5):159–162.
- Gleeson, P., Crook, S., Cannon, R. C., Hines, M. L., Billings, G. O., Farinella, M., Morse, T. M., Davison, A. P., Ray, S., Bhalla, U. S., Barnes, S. R., Dimitrova, Y. D., and Silver, R. A. (2010). NeuroML: A language for describing data driven models of neurons and networks with a high degree of biological detail. *PLoS Computational Biology*, 6(6):1–19.
- Hammer, M. and Champy, J. (1993). Reengineering the corporation: A manifesto for business revolution. *Business Horizons*, 36(5):90–91.
- Heiner, M. and Koch, I. (2004). Petri net based model validation in systems biology. In *Proceedings of the 25th International Conference on Application and Theory of Petri Nets*, pages 216–237. Springer.
- Henkel, R., Endler, L., Peters, A., Le Novère, N., and Waltemath, D. (2010). Ranked retrieval of computational biology models. *BMC bioinformatics*, 11(1):1.
- Henkel, R., Hoehndorf, R., Kacprowski, T., Knüpfer, C., Liebermeister, W., and Waltemath, D. (2016). Notions of similarity for systems biology models. *Briefings in Bioinformatics*, page bbw090.
- Henkel, R., Le Novère, N., Wolkenhauer, O., and Waltemath, D. (2012). Considerations of graph-based concepts to manage of computational biology models and associated simulations. In *Proceedings of the 2012 INFORMATIK conference: GI-Jahrestagung*, pages 1545–1551.
- Henkel, R., Wolkenhauer, O., and Waltemath, D. (2014). Combining computational models, semantic annotations, and associated simulation experiments in a graph database. *PeerJ*, pages 1–20.
- Hornung, T., Koschmider, A., and Lausen, G. (2008). Recommendation based process modeling support: Method and user experience. In *Proceedings of the 27th International Conference on Conceptual Modeling*, pages 265–278. Springer.

- Huber, S. (2015). A case mining based recommender system for knowledge workers. In *Proceedings of the Fifth International Symposium on Data-driven Process Discovery and Analysis*.
- Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., Arkin, A. P., Bornstein, B. J., Bray, D., Cornish-Bowden, A., Cuellar, A. A., Dronov, S., Gilles, E. D., Ginkel, M., Gor, V., Goryanin, I. I., Hedley, W. J., Hodgman, T. C., Hofmeyr, J. H., Hunter, P. J., Juty, N. S., Kasberger, J. L., Kremling, A., Kummer, U., Le Novère, N., Loew, L. M., Lucio, D., Mendes, P., Minch, E., Mjolsness, E. D., Nakayama, Y., Nelson, M. R., Nielsen, P. F., Sakurada, T., Schaff, J. C., Shapiro, B. E., Shimizu, T. S., Spence, H. D., Stelling, J., Takahashi, K., Tomita, M., Wagner, J., and Wang, J. (2003). The systems biology markup language (SBML): A medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531.
- Hucka, M., Hoops, S., Keating, S. M., Le Novère, N., Sahle, S., and Wilkinson, D. J. (2008). Systems biology markup language (sbml) level 2: Structures and facilities for model definitions. *Nature Precedings*.
- Hunger, M. (2014). *Neo4j 2.0: Eine Graphdatenbank für alle*. entwickler.press.
- Kim, M. and Leskovec, J. (2011). The network completion problem: Inferring missing nodes and edges in networks. In *Proceedings of the 2011 SIAM International Conference on Data Mining*, volume 11, pages 47–58. SIAM.
- Krause, F., Uhlenendorf, J., Lubitz, T., Schulz, M., Klipp, E., and Liebermeister, W. (2010). Annotation and merging of sbml models with semanticsbml. *Bioinformatics*, 26(3):421–422.
- Lambusch, F. (2015). *Towards Classifying Reactions of SBML-Models*. Bachelor thesis, Universität Rostock.
- Li, C., Donizelli, M., Rodriguez, N., Dharuri, H., Endler, L., Chelliah, V., Li, L., He, E., Henry, A., Stefan, M. I., Snoep, J. L., Hucka, M., Le Novère, N., and Laibe, C. (2010). BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC systems biology*, 4:92.
- Li, Y., Bandar, Z. A., and McLean, D. (2003). An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on knowledge and data engineering*, 15(4):871–882.
- Li, Y., Cao, B., Xu, L., Yin, J., Deng, S., Yin, Y., and Wu, Z. (2014). An efficient recommendation method for improving business process modeling. *IEEE Transactions on Industrial Informatics*, 10(1):502–513.
- Lloyd, C. M., Halstead, M. D. B., and Nielsen, P. F. (2004). CellML: Its future, present and past. *Progress in Biophysics and Molecular Biology*, 85(2):433–450.

- Maedche, A. and Staab, S. (2002). Measuring similarity between ontologies. In *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web*, pages 251–263. Springer.
- Melville, P. and Sindhvani, V. (2011). Recommender systems. In *Encyclopedia of machine learning*, pages 829–838. Springer.
- O’Donovan, J. and Smyth, B. (2005). Trust in recommender systems. In *Proceedings of the 10th international conference on Intelligent user interfaces*, pages 167–174. ACM.
- Partner, J. and Vukotic, A. (2012). *Neo4j in Action*. Manning Publications Co., meap edition.
- Peng, D., Steiniger, A., Helms, T., and Uhrmacher, A. M. (2013). Towards composing ml-rules models. In *Proceedings of the 2013 Winter Simulation Conference: Simulation: Making Decisions in a Complex World*, pages 4010–4011. IEEE Press.
- Pesquita, C., Faria, D., Falcao, A. O., Lord, P., and Couto, F. M. (2009). Semantic similarity in biomedical ontologies. *PLoS computational biology*, 5(7):e1000443.
- Pržulj, N. (2007). Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177–e183.
- Randhawa, R., Shaffer, C. A., and Tyson, J. (2010). Model composition for macromolecular regulatory networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 7(2):278–287.
- Randhawa, R., Shaffer, C. A., and Tyson, J. J. (2007). Fusing and composing macromolecular regulatory network models. In *Proceedings of the 2007 spring simulation multiconference-Volume 2*, pages 337–344. Society for Computer Simulation International.
- Randhawa, R., Shaffer, C. A., and Tyson, J. J. (2009). Model aggregation: a building-block approach to creating large macromolecular regulatory networks. *Bioinformatics*, 25(24):3289–3295.
- Ricci, F., Rokach, L., Shapira, B., and Kantor, P. B. (2011). *Recommender Systems Handbook*. Springer US.
- Robinson, I., Webber, J., and Eifrem, E. (2013). *Graph Databases*. O’Reilly Media, Inc.
- Rosenke, C. and Waltemath, D. (2014). How can semantic annotations support the identification of network similarities? In *Proceedings of the 7th International Workshop on Semantic Web Applications and Tools for Life Sciences*, volume 1320, page 11.
- Sadot, A., Fisher, J., Barak, D., Admanit, Y., Stern, M. J., Hubbard, E., and Harel, D. (2008). Toward verified biological models. *IEEE/ACM transactions on computational biology and bioinformatics*, 5(2):223–234.

- Schulz, M., Klipp, E., and Liebermeister, W. (2012). Propagating semantic information in biochemical network models. *BMC bioinformatics*, 13(1):1.
- Schulz, M., Krause, F., Le Novère, N., Klipp, E., and Liebermeister, W. (2011). Retrieval, alignment, and clustering of computational models based on semantic annotations. *Molecular systems biology*, 7(1):512.
- Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504.
- Smirnov, S., Weidlich, M., Mendling, J., and Weske, M. (2009). Action patterns in business process models. In *Proceedings of the 7th International Joint Conference on Service-Oriented Computing*, pages 115–129. Springer.
- Snoep, J. L. and Olivier, B. G. (2003). Jws online cellular systems modelling and microbiology. *Microbiology*, 149(11):3045–3047.
- Stanford, N. J., Wolstencroft, K., Golebiewski, M., Kania, R., Juty, N., Tomlinson, C., Owen, S., Butcher, S., Hermjakob, H., Le Novère, N., et al. (2015). The evolution of standards and data management practices in systems biology. *Molecular systems biology*, 11(12).
- Thavappiragasam, M., Lushbough, C. M., and Gnimpieba, E. Z. (2014). Automatic biosystems comparison using semantic and name similarity. In *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 790–796. ACM.
- Wang, J., Zhong, J., Chen, G., Li, M., Wu, F.-x., and Pan, Y. (2015). Clusterviz: a cytoscape app for cluster analysis of biological network. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 12(4):815–822.
- Wieloch, K., Filipowska, A., and Kaczmarek, M. (2011). Autocompletion for business process modelling. In *Proceedings of the International Conference on Business Information Systems*, pages 30–40. Springer.
- Yan, X. and Han, J. (2002). gSpan: graph-based substructure pattern mining. *Proceedings of the 2002 IEEE International Conference on Data Mining*, pages 721–724.
- Yang, Q. and Sze, S.-H. (2007). Path matching and graph matching in biological networks. *Journal of Computational Biology*, 14(1):56–67.

Appendix A: Survey

General results

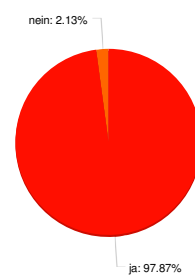
Support for modelling biological networks

1. Do you consider graphical tools to be helpful for the creation of biological networks? *

Anzahl Teilnehmer: 47

46 (97.9%): ja

1 (2.1%): nein

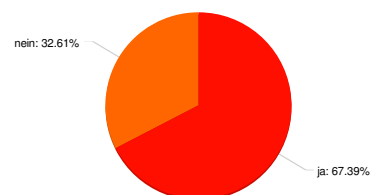


2. Do you personally use graphical tools to create biological networks? *

Anzahl Teilnehmer: 46

31 (67.4%): ja

15 (32.6%): nein



3. Which graphical tool(s) do you use to create biological networks? (Multiple choice)

Anzahl Teilnehmer: 29

11 (37.9%): CellDesigner

11 (37.9%): Cytoscape

1 (3.4%): SBGNViz

6 (20.7%): SBGN-ED

1 (3.4%): Wolfram
SystemModeler

1 (3.4%): BioUML

1 (3.4%): CARMEN

1 (3.4%): MonaLisa

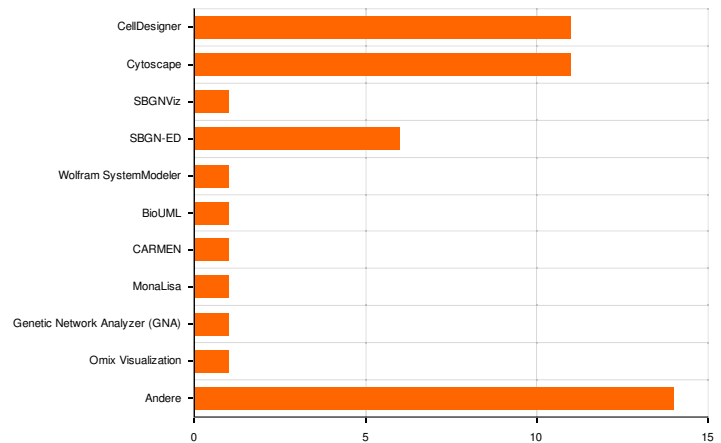
1 (3.4%): Genetic Network
Analyzer (GNA)

1 (3.4%): Omix Visualization

14 (48.3%): Andere

Antwort(en) aus dem Zusatzfeld:

- PowerPoint, Inkscape (just for visualizing)
- custom code
- r statistics
- ChiBE
- my own
- Escher (see <http://escher.github.io>)
- Jdesigner
- YeD
- Pathwaydesigner
- GINsim
- PathwayLab
- Inkscape, LibreOffice
- yEd
- Vanted

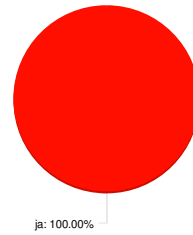


4. Do you consider the reuse of models or model networks helpful?

Anzahl Teilnehmer: 30

30 (100.0%): ja

- (0.0%): nein

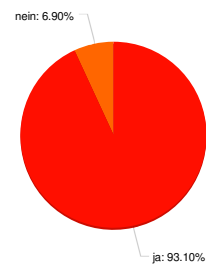


5. Do you think that a graphical tool can help you reuse models or model networks?

Anzahl Teilnehmer: 29

27 (93.1%): ja

2 (6.9%): nein

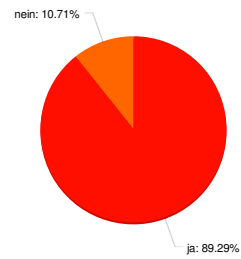


6. Would you consider it helpful to receive recommendations of similar models to compare them with your network?

Anzahl Teilnehmer: 28

25 (89.3%): ja

3 (10.7%): nein

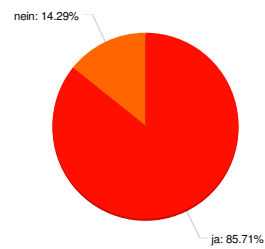


7. Would you consider it helpful to receive recommendations of subnetworks to extend your network?

Anzahl Teilnehmer: 28

24 (85.7%): ja

4 (14.3%): nein

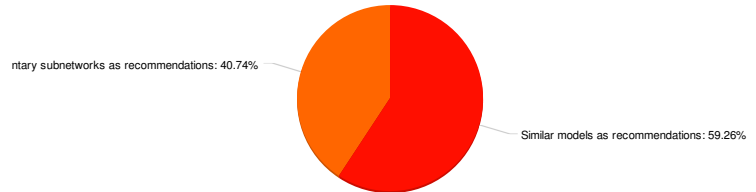


8. Would you prefer to compare your network to similar models, or to receive recommended subnetworks to extend your model?

Anzahl Teilnehmer: 27

16 (59.3%): Similar models as recommendations

11 (40.7%): Complementary subnetworks as recommendations

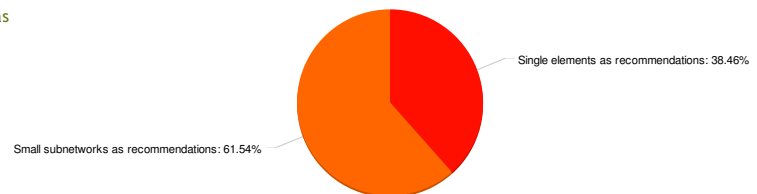


9. Would you prefer to receive recommendations of single elements or of subnetworks containing more than one element to extend your created network?

Anzahl Teilnehmer: 26

10 (38.5%): Single elements as recommendations

16 (61.5%): Small subnetworks as recommendations

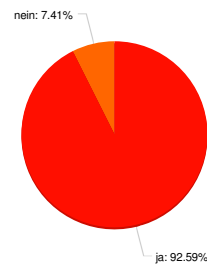


10. Do you consider recommendations on the basis of a repository (e.g. BioModels, Physiome Model Repository) helpful?

Anzahl Teilnehmer: 27

25 (92.6%): ja

2 (7.4%): nein

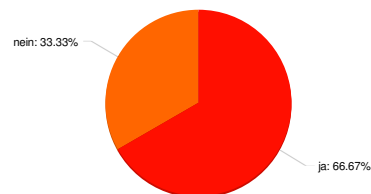


11. Do you consider recommendations on the basis of a selected smaller set of models (e.g. only cell cycle models) helpful?

Anzahl Teilnehmer: 27

18 (66.7%): ja

9 (33.3%): nein

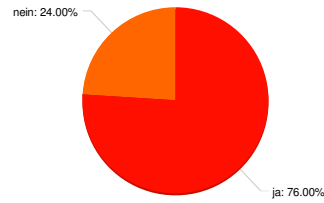


12. Do you consider it helpful to base the recommendations on your own customised set of models?

Anzahl Teilnehmer: 25

19 (76.0%): ja

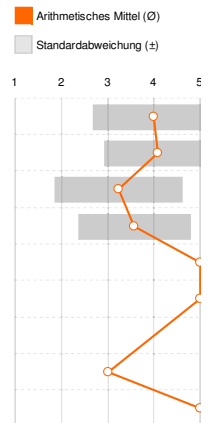
6 (24.0%): nein



13. How important do you consider the following information when selecting appropriate subnetworks from a list of recommended subnetworks?

Anzahl Teilnehmer: 26

	1 (unimportant) (1)		2 (2)		3 (3)		4 (4)		5 (important) (5)		don't know (0)		
	Σ	%	Σ	%	Σ	%	Σ	%	Σ	%	Σ	Ø	±
Image of the subnetwork	1x	3,85	4x	15,38	4x	15,38	2x	7,69	15x	57,69	-	4,00	1,33
Names of subnetwork ele...	1x	3,85	2x	7,69	4x	15,38	6x	23,08	13x	50,00	-	4,08	1,16
Ontology terms of eleme...	3x	11,54	6x	23,08	6x	23,08	4x	15,38	7x	26,92	-	3,23	1,39
Information regarding th...	1x	3,85	5x	19,23	5x	19,23	8x	30,77	7x	26,92	-	3,58	1,21
Biological classifiers (ce...	-	-	-	-	-	-	-	-	1x	100,00	-	5,00	0,00
experimental details	-	-	-	-	-	-	-	-	1x	100,00	-	5,00	0,00
glitch	-	-	-	-	-	-	-	-	-	-	1x	-	-
In a disease context, how...	-	-	-	-	1x	100,00	-	-	-	-	-	3,00	0,00
Technical classifiers (pa...	-	-	-	-	-	-	-	-	1x	100,00	-	5,00	0,00

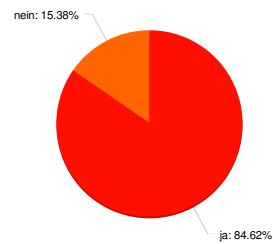


14. Do you consider it helpful to have filters to limit the number of recommendations?

Anzahl Teilnehmer: 26

22 (84.6%): ja

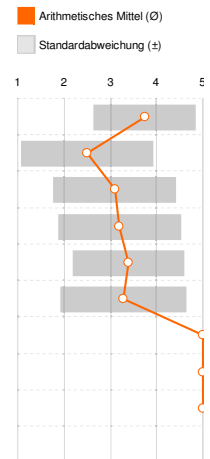
4 (15.4%): nein



15. Filtering options: how important do you consider the following options?

Anzahl Teilnehmer: 21

	1 (unimportant) (1)		2 (2)		3 (3)		4 (4)		5 (important) (5)		don't know (0)		
	Σ	%	Σ	%	Σ	%	Σ	%	Σ	%	Σ	Ø	±
Maximum number of reco...	1x	5,00	1x	5,00	5x	25,00	7x	35,00	5x	25,00	1x	3,74	1,10
Minimum size of recomm...	6x	28,57	6x	28,57	3x	14,29	2x	9,52	3x	14,29	1x	2,50	1,43
Maximum size of recomm...	2x	9,52	6x	28,57	4x	19,05	4x	19,05	4x	19,05	1x	3,10	1,33
Ontology type (e.g. Gene..	2x	9,52	5x	23,81	4x	19,05	5x	23,81	4x	19,05	1x	3,20	1,32
Ontology level (abstract...	1x	4,76	3x	14,29	6x	28,57	4x	19,05	4x	19,05	3x	3,39	1,20
Number of occurrences o...	3x	15,00	1x	5,00	6x	30,00	4x	20,00	4x	20,00	2x	3,28	1,36
Quality of model	-	-	-	-	-	-	-	-	1x	100,00	-	5,00	0,00
Statistics (see above. Se...	-	-	-	-	-	-	-	-	1x	100,00	-	5,00	0,00
Technical and biological...	-	-	-	-	-	-	-	-	1x	100,00	-	5,00	0,00
this seems specific to w...	-	-	-	-	-	-	-	-	-	-	1x	-	-



CellDesigner users only

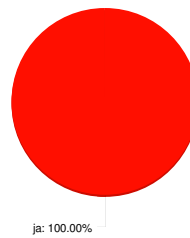
Support for modelling biological networks

1. Do you consider graphical tools to be helpful for the creation of biological networks? *

Anzahl Teilnehmer: 11

11 (100.0%): ja

- (0.0%): nein

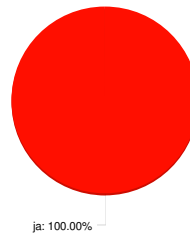


2. Do you personally use graphical tools to create biological networks? *

Anzahl Teilnehmer: 11


11 (100.0%): ja

- (0.0%): nein



3. Which graphical tool(s) do you use to create biological networks? (Multiple choice)

Anzahl Teilnehmer: 11

11 (100.0%): CellDesigner 

5 (45.5%): Cytoscape

- (0.0%): SBGNViz

2 (18.2%): SBGN-ED

1 (9.1%): Wolfram
SystemModeler

- (0.0%): Ben(Zai)Ten

- (0.0%): Athena

- (0.0%): BioUML

1 (9.1%): CARMEN

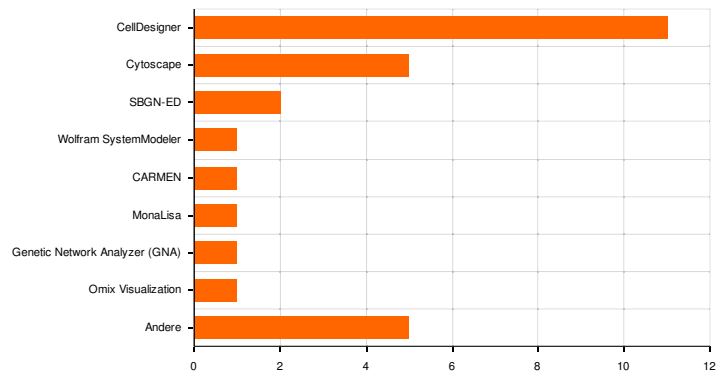
1 (9.1%): MonaLisa

1 (9.1%): Genetic Network
Analyzer (GNA)

1 (9.1%): Omix Visualization

- (0.0%): SimBiology

5 (45.5%): Andere



Antwort(en) aus dem Zusatzfeld:

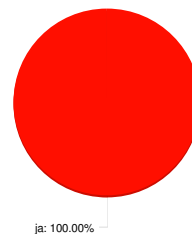
- PowerPoint, Inkscape (just for visualizing)
- Escher (see <http://escher.github.io>)
- YeD
- PathwayLab
- Inkscape, LibreOffice

4. Do you consider the reuse of models or model networks helpful?

Anzahl Teilnehmer: 11

11 (100.0%): ja

- (0.0%): nein

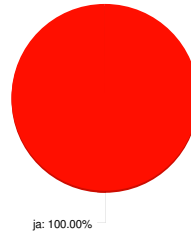


5. Do you think that a graphical tool can help you reuse models or model networks?

Anzahl Teilnehmer: 11

11 (100.0%): ja

- (0.0%): nein

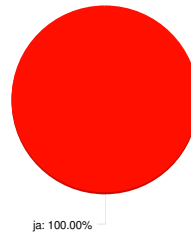


6. Would you consider it helpful to receive recommendations of similar models to compare them with your network?

Anzahl Teilnehmer: 11

11 (100.0%): ja

- (0.0%): nein

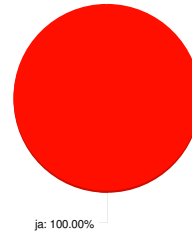


7. Would you consider it helpful to receive recommendations of subnetworks to extend your network?

Anzahl Teilnehmer: 11

11 (100.0%): ja

- (0.0%): nein

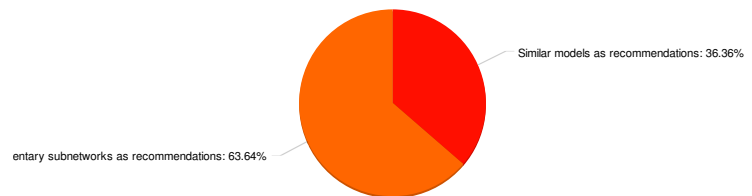


8. Would you prefer to compare your network to similar models, or to receive recommended subnetworks to extend your model?

Anzahl Teilnehmer: 11

4 (36.4%): Similar models as recommendations

7 (63.6%): Complementary subnetworks as recommendations

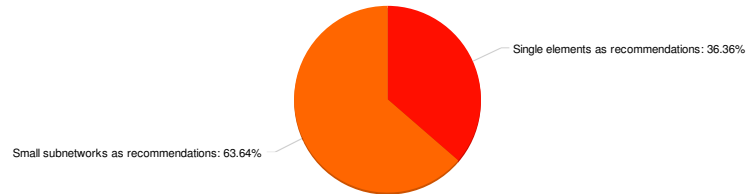


9. Would you prefer to receive recommendations of single elements or of subnetworks containing more than one element to extend your created network?

Anzahl Teilnehmer: 11

4 (36.4%): Single elements as recommendations

7 (63.6%): Small subnetworks as recommendations

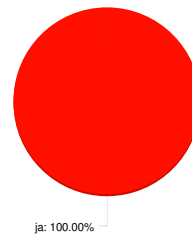


10. Do you consider recommendations on the basis of a repository (e.g. BioModels, Physiome Model Repository) helpful?

Anzahl Teilnehmer: 11

11 (100.0%): ja

- (0.0%): nein

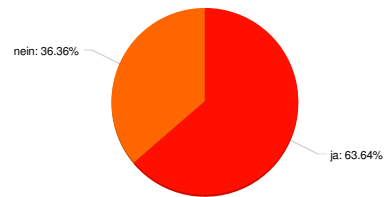


11. Do you consider recommendations on the basis of a selected smaller set of models (e.g. only cell cycle models) helpful?

Anzahl Teilnehmer: 11

7 (63.6%): ja

4 (36.4%): nein

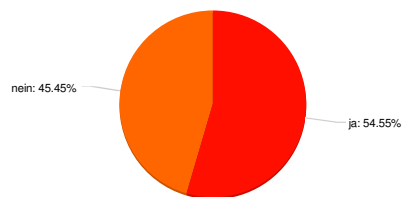


12. Do you consider it helpful to base the recommendations on your own customised set of models?

Anzahl Teilnehmer: 11

6 (54.5%): ja

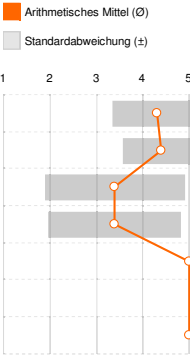
5 (45.5%): nein



13. How important do you consider the following information when selecting appropriate subnetworks from a list of recommended subnetworks?

Anzahl Teilnehmer: 10

	1 (unimportant) (1)		2 (2)		3 (3)		4 (4)		5 (important) know (5)		don't know (0)	
	Σ	%	Σ	%	Σ	%	Σ	%	Σ	%	Σ	Ø ±
Image of the subnetwork	-	-	-	-	3x	30,00	1x	10,00	6x	60,00	-	4,30 0,95
Names of subnetwork ele..	-	-	-	-	2x	20,00	2x	20,00	6x	60,00	-	4,40 0,84
Ontology terms of eleme..	1x	10,00	3x	30,00	-	-	3x	30,00	3x	30,00	-	3,40 1,51
Information regarding th...	1x	10,00	2x	20,00	2x	20,00	2x	20,00	3x	30,00	-	3,40 1,43
Biological classifiers (ce...	-	-	-	-	-	-	-	-	1x	100,00	-	5,00 0,00
glitch	-	-	-	-	-	-	-	-	-	-	1x	- -
Technical classifiers (pa...	-	-	-	-	-	-	-	-	1x	100,00	-	5,00 0,00

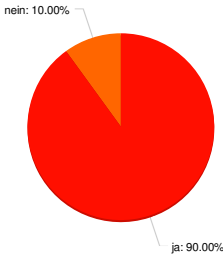


14. Do you consider it helpful to have filters to limit the number of recommendations?

Anzahl Teilnehmer: 10

9 (90.0%): ja

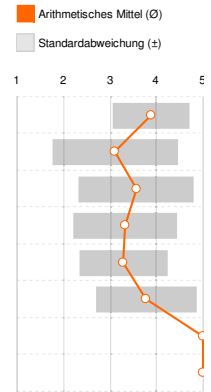
1 (10.0%): nein



15. Filtering options: how important do you consider the following options?

Anzahl Teilnehmer: 9

	1 (unimportant) (1)		2 (2)		3 (3)		4 (4)		5 (important) know (5)		don't know (0)	
	Σ	%	Σ	%	Σ	%	Σ	%	Σ	%	Σ	Ø ±
Maximum number of reco..	-	-	-	-	3x	37,50	3x	37,50	2x	25,00	-	3,88 0,83
Minimum size of recomm...	1x	11,11	2x	22,22	3x	33,33	1x	11,11	2x	22,22	-	3,11 1,36
Maximum size of recomm..	-	-	2x	22,22	3x	33,33	1x	11,11	3x	33,33	-	3,56 1,24
Ontology type (e.g. Gene..	-	-	2x	22,22	4x	44,44	1x	11,11	2x	22,22	-	3,33 1,12
Ontology level (abstract...	-	-	1x	11,11	4x	44,44	1x	11,11	1x	11,11	2x	3,29 0,95
Number of occurrences o..	-	-	1x	11,11	3x	33,33	2x	22,22	3x	33,33	-	3,78 1,09
Statistics (see above. Se..	-	-	-	-	-	-	-	-	1x	100,00	-	5,00 0,00
Technical and biological...	-	-	-	-	-	-	-	-	1x	100,00	-	5,00 0,00



Cytoscape users only

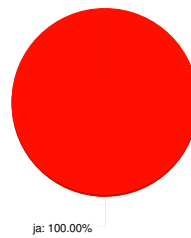
Support for modelling biological networks

1. Do you consider graphical tools to be helpful for the creation of biological networks? *

Anzahl Teilnehmer: 11

11 (100.0%): ja

- (0.0%): nein

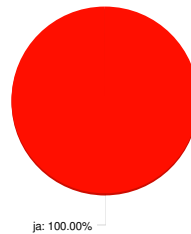


2. Do you personally use graphical tools to create biological networks? *

Anzahl Teilnehmer: 11

11 (100.0%): ja


- (0.0%): nein



3. Which graphical tool(s) do you use to create biological networks? (Multiple choice)

Anzahl Teilnehmer: 11

5 (45.5%): CellDesigner

11 (100.0%): Cytoscape 

- (0.0%): SBGNViz

1 (9.1%): SBGN-ED

- (0.0%): Wolfram
SystemModeler

- (0.0%): Ben(Zai)Ten

- (0.0%): Athena

- (0.0%): BioUML

- (0.0%): CARMEN

- (0.0%): MonaLisa

1 (9.1%): Genetic Network
Analyzer (GNA)

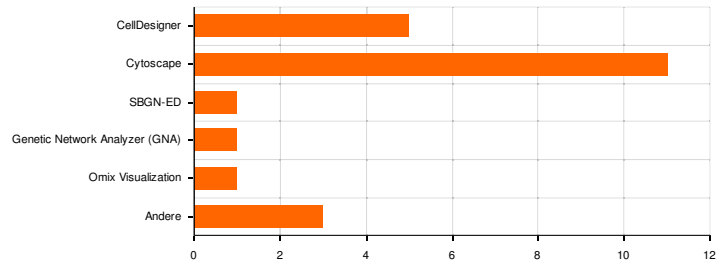
1 (9.1%): Omix Visualization

- (0.0%): SimBiology

3 (27.3%): Andere

Antwort(en) aus dem Zusatzfeld:

- PowerPoint, Inkscape (just for
visualizing)
- Escher (see
<http://escher.github.io>)
- Inkscape, LibreOffice

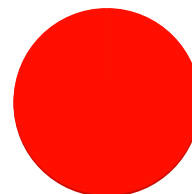


4. Do you consider the reuse of models or model networks helpful?

Anzahl Teilnehmer: 11

11 (100.0%): ja

- (0.0%): nein



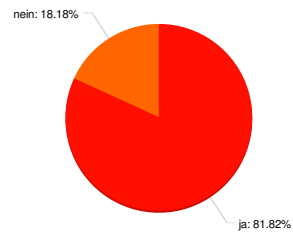
ja: 100.00%

5. Do you think that a graphical tool can help you reuse models or model networks?

Anzahl Teilnehmer: 11

9 (81.8%): ja

2 (18.2%): nein

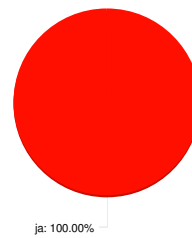


6. Would you consider it helpful to receive recommendations of similar models to compare them with your network?

Anzahl Teilnehmer: 11

11 (100.0%): ja

- (0.0%): nein

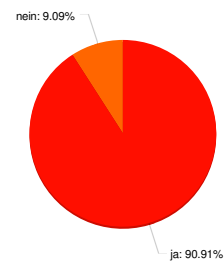


7. Would you consider it helpful to receive recommendations of subnetworks to extend your network?

Anzahl Teilnehmer: 11

10 (90.9%): ja

1 (9.1%): nein

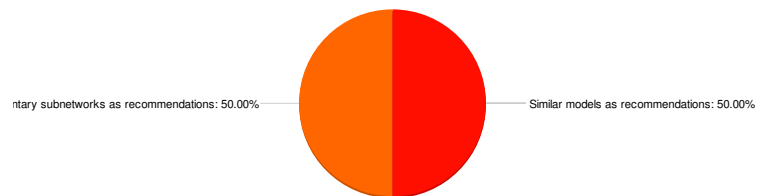


8. Would you prefer to compare your network to similar models, or to receive recommended subnetworks to extend your model?

Anzahl Teilnehmer: 10

5 (50.0%): Similar models as recommendations

5 (50.0%): Complementary subnetworks as recommendations

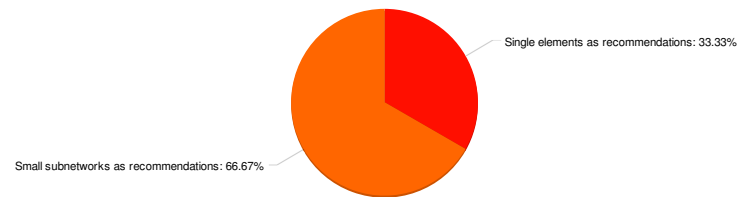


9. Would you prefer to receive recommendations of single elements or of subnetworks containing more than one element to extend your created network?

Anzahl Teilnehmer: 9

3 (33.3%): Single elements as recommendations

6 (66.7%): Small subnetworks as recommendations

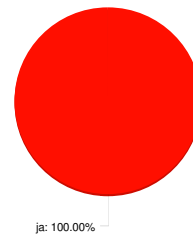


10. Do you consider recommendations on the basis of a repository (e.g. BioModels, Physiome Model Repository) helpful?

Anzahl Teilnehmer: 10

10 (100.0%): ja

- (0.0%): nein

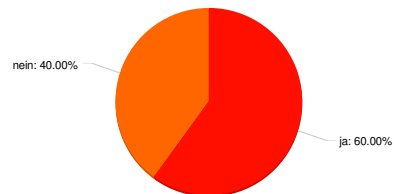


11. Do you consider recommendations on the basis of a selected smaller set of models (e.g. only cell cycle models) helpful?

Anzahl Teilnehmer: 10

6 (60.0%): ja

4 (40.0%): nein

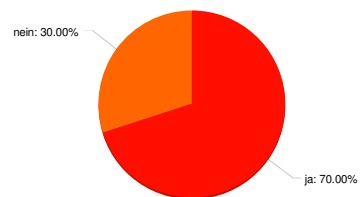


12. Do you consider it helpful to base the recommendations on your own customised set of models?

Anzahl Teilnehmer: 10

7 (70.0%): ja

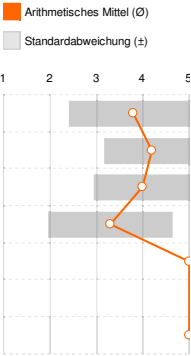
3 (30.0%): nein



13. How important do you consider the following information when selecting appropriate subnetworks from a list of recommended subnetworks?

Anzahl Teilnehmer: 10

	1 (unimportant) (1)		2 (2)		3 (3)		4 (4)		5 (important) know (5)		don't know (0)	
	Σ	%	Σ	%	Σ	%	Σ	%	Σ	%	Σ	Ø ±
Image of the subnetwork	-	-	3x	30,00	1x	10,00	1x	10,00	5x	50,00	-	3,80 1,40
Names of subnetwork ele..	-	-	1x	10,00	1x	10,00	3x	30,00	5x	50,00	-	4,20 1,03
Ontology terms of eleme..	-	-	1x	10,00	2x	20,00	3x	30,00	4x	40,00	-	4,00 1,05
Information regarding th...	1x	10,00	2x	20,00	2x	20,00	3x	30,00	2x	20,00	-	3,30 1,34
Biological classifiers (ce...	-	-	-	-	-	-	-	-	1x	100,00	-	5,00 0,00
glitch	-	-	-	-	-	-	-	-	-	-	1x	- -
Technical classifiers (pa...	-	-	-	-	-	-	-	-	1x	100,00	-	5,00 0,00

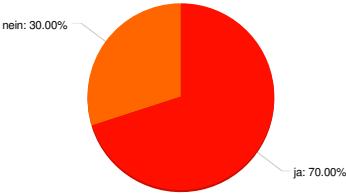


14. Do you consider it helpful to have filters to limit the number of recommendations?

Anzahl Teilnehmer: 10

7 (70.0%): ja

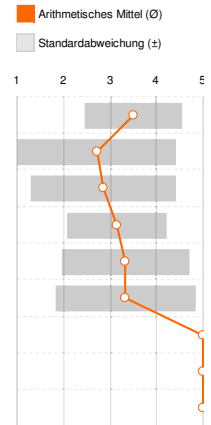
3 (30.0%): nein



15. Filtering options: how important do you consider the following options?

Anzahl Teilnehmer: 7

	1 (unimportant) (1)		2 (2)		3 (3)		4 (4)		5 (important) know (5)		don't know (0)	
	Σ	%	Σ	%	Σ	%	Σ	%	Σ	%	Σ	Ø ±
Maximum number of reco..	-	-	1x	16,67	2x	33,33	2x	33,33	1x	16,67	-	3,50 1,05
Minimum size of recomm...	2x	28,57	2x	28,57	1x	14,29	-	-	2x	28,57	-	2,71 1,70
Maximum size of recomm..	1x	14,29	3x	42,86	1x	14,29	-	-	2x	28,57	-	2,86 1,57
Ontology type (e.g. Gene..	1x	14,29	-	-	3x	42,86	3x	42,86	-	-	-	3,14 1,07
Ontology level (abstract...	1x	14,29	-	-	2x	28,57	2x	28,57	1x	14,29	1x	3,33 1,37
Number of occurrences o..	1x	16,67	-	-	3x	50,00	-	-	2x	33,33	-	3,33 1,51
Quality of model	-	-	-	-	-	-	-	-	1x	100,00	-	5,00 0,00
Statistics (see above. Se..	-	-	-	-	-	-	-	-	1x	100,00	-	5,00 0,00
Technical and biological...	-	-	-	-	-	-	-	-	1x	100,00	-	5,00 0,00



Appendix B: SBML models

Exemplary SBML file

Listing 2: SBML structure of the model named “Tyson1991 - Cell Cycle 6 var” exported from BioModels (ID BIOMD0000000005)

```
<?xml version='1.0' encoding='UTF-8' standalone='no'?>
<sbml xmlns="http://www.sbml.org/sbml/level2/version4" ...>
  <model id="BIOMD0000000005" name="Tyson1991 - Cell Cycle 6 var" ... >
    <notes> ... </notes> <annotation> ... </annotation>
    <listOfCompartments> ... </listOfCompartments>
    <listOfSpecies>
      <species id="C2" initialAmount="0" name="cdc2k" ... >
        <annotation>
          <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
            xmlns:bqmodel="http://biomodels.net/model-qualifiers/"
            xmlns:bqbiol="http://biomodels.net/biology-qualifiers/">
            <rdf:Description rdf:about="#_000004">
              <bqbiol:isVersionOf>
                <rdf:Bag>
                  <rdf:li rdf:resource="http://identifiers.org/uniprot/P04551"/>
                </rdf:Bag>
              </bqbiol:isVersionOf>
            </rdf:Description>
          </rdf:RDF>
        </annotation>
      </species>
      ...
    </listOfSpecies>
    <listOfRules>
      ...
    </listOfRules>
    <listOfReactions>
      <reaction id="Reaction2" name="cdc2k phosphorylation" ... >
        <annotation>
          <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
            xmlns:bqmodel="http://biomodels.net/model-qualifiers/"
            xmlns:bqbiol="http://biomodels.net/biology-qualifiers/">
            <rdf:Description rdf:about="#_000011">
              <bqbiol:isVersionOf>
                <rdf:Bag>
                  <rdf:li rdf:resource="http://identifiers.org/ec-code/2.7.11.1"/>
                  <rdf:li rdf:resource="http://identifiers.org/obo.go/G0:0006468"/>
                </rdf:Bag>
              </bqbiol:isVersionOf>
            </rdf:Description>
          </rdf:RDF>
        </annotation>
        <listOfReactants>
          <speciesReference species="C2" metaid="_712418"/>
        </listOfReactants>
        <listOfProducts>
          <speciesReference species="CP" metaid="_712430"/>
        </listOfProducts>
        <kineticLaw metaid="_712442"> ... </kineticLaw>
      </reaction>
      ...
    </listOfReactions>
  </model>
</sbml>
```

Appendix C: Exemplary Cypher queries for strategies

Strategy 1

Listing 3: Exemplary Cypher query to receive direct neighbours of a selected entity

```
MATCH
    (reac:SBML_REACTION)-[:HAS_ANNOTATION]->(ar:ANNOTATION),
    (ar)-[:isVersionOf]->(r1:RESOURCE),
    (ar)-[:isVersionOf]->(r2:RESOURCE),
    (reac)-[rel]->(s:SBML_SPECIES),
    (s)-[:HAS_ANNOTATION]->(as:ANNOTATION)-[qualifier]->(rs:RESOURCE),
    (m:SBML_MODEL)-[:HAS_REACTION]->(reac)
WHERE
    r1.URI=~".*2.7.11.1" OR r2.URI=~".*GO:0006468"
RETURN
    TYPE(rel) as edgeType, labels(s) as nodeType,
    TYPE(qualifier) as qualifier, rs.URI as URI,
    COUNT(DISTINCT ID(m)) as numberOfModels,
    COUNT(DISTINCT s) as frequency
ORDER BY numberOfModels DESC, frequency DESC
```

Strategy 2.1

Listing 4: Exemplary Cypher query to receive direct neighbours of a selected entity based on the whole structure of the model under construction

```
MATCH
    (cdc2k:SBML_SPECIES)-[:IS_REACTANT]->(phos:SBML_REACTION),
    (phos)-[:HAS_PRODUCT]->(cdc2kp:SBML_SPECIES),
    (cdc2kp)-[:IS_REACTANT]->(dephos:SBML_REACTION),
    (dephos)-[:HAS_PRODUCT]->(cdc2k),
    (cdc2k)-[rel]->(reac:SBML_REACTION),
    (reac)-[:HAS_ANNOTATION]->(a:ANNOTATION)-[qualifier]->(r:RESOURCE),
    (m:SBML_MODEL)-[:HAS_REACTION]->(reac)
RETURN
    TYPE(rel) as edgeType, labels(reac) as nodeType,
    TYPE(qualifier) as qualifier, r.URI as URI,
    COUNT(DISTINCT ID(m)) as numberOfModels,
    COUNT(DISTINCT cdc2k) as frequency
ORDER BY numberOfModels DESC, frequency DESC
```

Strategy 2.2

Listing 5: Exemplary Cypher query to receive direct neighbours of a selected entity based on the whole model under construction with annotations

```
MATCH
    (cdc2k:SBML_SPECIES)-[:IS_REACTANT]->(phos:SBML_REACTION),
    (cdc2k)-[:HAS_ANNOTATION]->(annCdc2k:ANNOTATION),
    (annCdc2k)-[:isVersionOf]->(resCdc2k:RESOURCE),
    (phos)-[:HAS_ANNOTATION]->(annPhos:ANNOTATION),
    (annPhos)-[:isVersionOf]->(res1phos:RESOURCE),
    (annPhos)-[:isVersionOf]->(res2phos:RESOURCE),
    (cdc2k)-[rel]->(reac:SBML_REACTION),
    (reac)-[:HAS_ANNOTATION]->(:ANNOTATION)-[qualifier]->(r:RESOURCE),
    (m:SBML_MODEL)-[:HAS_REACTION]->(reac)
WHERE
    resCdc2k.URI=~".*P04551" AND
    res1phos.URI=~".*2.7.11.1" AND
    res2phos.URI=~".*G0:0006468"
RETURN
    TYPE(rel) as edgeType, labels(reac) as nodeType,
    TYPE(qualifier) as qualifier, r.URI as URI,
    COUNT(DISTINCT ID(m)) as numberOfModels,
    COUNT(DISTINCT cdc2k) as frequency
    Order BY numberOfModels DESC, frequency DESC
```

Strategy 4

Listing 6: Exemplary Cypher query to receive semantically similar models

```
MATCH
    (m:SBML_MODEL)-[:HAS_SPECIES|HAS_REACTION]->(entity),
    (entity)-[:HAS_ANNOTATION]->(:ANNOTATION)-[qualifier]->(r:RESOURCE)
WHERE
    r.URI=~'.*uniprot/P04551' OR
    r.URI=~'.*ec-code/2.7.11.1' OR
    r.URI=~'.*G0:0006468'
RETURN
    ID(m) AS internalID, m.NAME as modelName,
    collect(entity.NAME) as matchingEntityNames,
    collect(DISTINCT r.URI) as matchingURIs,
    COUNT(DISTINCT entity) as numEntities,
    COUNT(DISTINCT r.URI) as numURIs,
    toFloat(COUNT(DISTINCT r.URI)/(COUNT(DISTINCT entity))) as proportion
    Order BY numURIs DESC, proportion DESC
```

Strategy 6.2

Listing 7: Exemplary Cypher query to receive entities reachable in three steps

```
MATCH
    (m:SBML_MODEL)-[:HAS_SPECIES|HAS_REACTION]->(entity),
    (entity)-[:HAS_ANNOTATION]->(:ANNOTATION)-[]->(r:RESOURCE)
WHERE
    (ID(m)=102286 OR ID(m)=105835 OR ID(m)=104196 OR
     ID(m)=122372 OR ID(m)=105090) AND
    (r.URI=~'.*uniprot/P04551' OR
     r.URI=~'.*ec-code/2.7.11.1' OR
     r.URI=~'.*GO:0006468')
WITH
    entity,m
MATCH
    (entity)-[rel:HAS_REACTANT|IS_REACTANT|HAS_PRODUCT|
               IS_PRODUCT|HAS_MODIFIER|IS_MODIFIER*1..3]->(neighbour),
    (neighbour)-[:HAS_ANNOTATION]->(:ANNOTATION)-[qualifier]->(rN:RESOURCE)
RETURN
    labels(neighbour) as nodeType, TYPE(qualifier) as qualifier,
    rN.URI as URI, COUNT(DISTINCT rel) as freqRelated,
    COUNT(DISTINCT m) as numOfModels
ORDER BY freqRelated DESC, numOfModels DESC
```