Pathway Computation in Models Derived from Bio-science Text Sources

Andreasen, Troels; Styltsvig, Henrik Bulskov; Jensen, Per Anker; Nilsson, Jørgen Fischer

Document Version
Accepted author manuscript

Published in:
Foundations of Intelligent Systems - 23rd International Symposium, ISMIS 2017, Proceedings

DOI:
10.1007/978-3-319-60438-1_42

Publication date:
2017

License
Unspecified

Citation for published version (APA):
Andreasen, T., Styltsvig, H. B., Jensen, P. A., & Nilsson, J. F. (2017). Pathway Computation in Models Derived from Bio-science Text Sources. In M. Kryszkiewicz , A. Appice, D. Ślęzak, H. Rybinski, A. Skowron, & Z. W. Raś (Eds.), *Foundations of Intelligent Systems - 23rd International Symposium, ISMIS 2017, Proceedings: Proceedings of the 23rd International Symposium, ISMIS 2017* (Vol. 10352, pp. 424-434). Springer. https://doi.org/10.1007/978-3-319-60438-1_42

Pathway Computation in Models Derived from Bio-science Text Sources

Troels Andreasen, Henrik Bulskov Styltsvig, Per Anker Jensen, and Jørgen Fischer Nilsson

Article in proceedings (Accepted version*)

Pathway Computation in Models Derived from Bio-Science Text Sources

Troels Andreasen[1], Henrik Bulskov[1], Per Anker Jensen[2], Jørgen Fischer Nilsson[3]

[1]Computer Science, Roskilde University,
[2]Management, Society and Communication, Copenhagen Business School
[3]Mathematics and Computer Science, Technical University of Denmark,
{troels,bulskov}@ruc.dk, paj.msc@cbs.dk, jfni@dtu.dk

Abstract. This paper outlines a system, OntoScape, serving to accomplish complex inference tasks on knowledge bases and bio-models derived from life-science text corpora. The system applies so-called natural logic, a form of logic which is readable for humans. This logic affords ontological representations of complex terms appearing in the text sources. Along with logical propositions, the system applies a semantic graph representation facilitating calculation of bio-pathways. More generally, the system affords means of query answering appealing to general and domain specific inference rules.
Keywords: Semantic text processing in bio-informatics; bio-models using natural logic and semantic graphs; querying and pathway computation

1 Introduction

This paper addresses logic-based bio-models derived from life science texts. We discuss representation languages and reasoning principles for bio-models derived from actual life science sources. In particular, we describe and exemplify the intended query answering and pathway functionality, that is, the ability to compute conceptual pathways in the stored model.

Our approach is based on the construction of a logical model for a considered bio-system and is in line with the foundational developments in [1, 2]. One main challenge in the logical approach is the extraction of comprehensive bio-models from text sources and formalisation of these. This logical approach is in contrast to established and rather succesful approaches to text mining based on direct references to phrases in concrete text sources and advanced information extraction techniques, cf. for example [13–16].

At first sight our approach resembles the well-known, rather simplistic entity-relationship models and RDF representations. However, our framework is unique in various respects, first of all in its generativity, that is, the ability of the models to accommodate arbitrarily complex concepts formed by composition of lexicalized classes and relationships as discussed in [3]. By way of examples, the virtually open-ended supply of concept terms such as *'cell in the liver that secretes hormone'*, *'arteria in pancreas'*, *'secrete from the exocrine pancreas'* are

accomodated in the model by composing simple, given class terms into compound concept terms with an obvious resemblance to phrases in natural language, as these examples illustrate. All such encountered concepts are situated in the so-called "generative ontology" in a manner such that they can be "de-constructed" and reasoned with computationally.

The generativity and the liaison to natural language specifications is achieved, as mentioned, by adopting 'natural logic' cf. [6, 7] as the logical model language. In addition, the models come in the form of graphs with concepts as nodes and relations as edges.

The paper is organized as follows: In section 2 we introduce models formalized in terms of natural logic. In section 3 we derive semantic graph-based models from the natural logic specifications and in section 4 we exemplify natural logic model fragments drawn from various medical text sources. Section 5 describes our prototype system and explains the pathway computation functionality, concluding finally in section 6.

2 Models in Natural Logic

In the applied natural logic conception, a knowledge base or specification consists of a collection of descriptive sentences called 'propositions' in order to distinguish them from the natural language sentences from which they are derived. Propositions in the applied logic, dubbed NaturaLog, are of the following general form

$$Cterm_1\ Relterm\ Cterm_2$$

where

- The two *Cterm*s are atomic or compound concept terms.
- The *Relterm* is a relational term, in the simplest cases corresponding to a transitive verb, e.g. '*cause*', or '*secrete*' or prepositions like '*in*', '*via*' etc.

In a logical proposition like betacell secrete insulin the two concept terms are atomic, and so is the intervening relational term. A bio-model comprising also the class inclusion ontology consists of a finite, albeit possibly huge, collection of such NaturaLog propositions. As it appears, we use *sans serif* font for propositions throughout. This model is then the basis for inferences and querying.

Propositions may contain complex structures: Compound *Cterm*s consist of a class C with attached qualifications. In a more complex proposition like

(cell that secrete insulin) is:located:in (pancreatic gland).

the first concept term consists of the atomic term cell adorned with a relative clause consisting of the relational term secrete followed by the concept term insulin. Relative clauses are indicated by the optional keyword 'that', merely to make the reading easier. Relative clauses are assumed always to act restrictively. For instance, as a matter of principle, cell that secrete insulin is recorded by the system as a sub-concept of cell in the concept inclusion structure in the ontology. Likewise, the second concept term pancreatic gland, is recognized as a sub-concept of the class gland in that all adjectives are also assumed to be

interpreted restrictively. Parentheses are inserted for ease of reading and serve to ensure disambiguation. They may be omitted if there is no risk of ambiguity.

However, sub-class - and, more generally, sub-concept relationships may also be specified explicitly, namely by the relation term isa, as seen in copula sentences. Example: betacell isa cell. By contrast, the propositions (cell that secretes insulin) isa cell and (pancreatic cell) isa cell are inferred by the system according to the principles mentioned. Still, (pancreatic cell) isa (cell located-in pancreas) (and *vice versa*) has to be provided.

As it appears, the natural logic propositions are perfectly readable, if somewhat stereotypical, by domain experts by virtue of their resemblance to natural language. The converse, challenging task of automating translation from manageable parts of natural language in scientific text sources into natural logic is approached in our [3].

2.1 Quantifiers and Recursion in Concept Terms

The above propositional form $Cterm_1$ $Relterm$ $Cterm_2$ is a special case of

$$Q_1\ Cterm_1\ Relterm\ Q_2\ Cterm_2$$

where the Qs are quantifiers, primarily 'all/every' or 'some'. Usually the quantifiers are absent with Q_1 then being interpreted as all and Q_2 as some by default. Accordingly, the example betacell secrete insulin is interpreted logically as the proposition all betacell secrete some insulin, where some insulin is meant to be some portion or amount of insulin. Generally speaking, classes are assumed to be non-empty (appealing to existential import), and the entities in a class of substance are taken to be arbitrary, non-empty amounts of the substance.

The propositional form all $Cterm_1$ $Relterm$ some $Cterm_2$ corresponds to the predicate logic formula $\forall x(Cterm_1[x] \rightarrow \exists y(Relterm[x, y] \wedge Cterm_2[y]))$, see further [3, 4], where we also discuss the relationship to description logic. The introduced NaturaLog forms cover only those parts of binary predicate logic which are considered relevant for bio-modelling. Notable exclusions at present are logical negation and logical disjunctions.

Recall that a concept term consists of a class C followed by one or more qualifications or restrictions, where restrictions consist of a relational term followed by a concept term: $Relterm$ $Cterm$. In case of more than one restriction, these are to form a conjunction with and understood as logical conjunction proper. By contrast, two and-aligned concept terms within the same class are conceived of as a logical disjunction (ex. beta-cell and alpha-cell produce hormone). Accordingly, concept terms have a finitely nested, recursive structure reflecting the syntax of natural language nominal phrases with possibly nested relative clauses and prepositional phrases. The handling of adjectives (ex. pancreatic gland) and compound nouns (ex. lung symptom) are both assumed to be acting restrictively. These as well as genitives will not be discussed further in this paper.

2.2 Ontologies

As mentioned above, a special case of the above propositions is class inclusion relationships corresponding to stylized copula sentences. For example, in the proposition pancreas isa (endocrine gland), isa denotes concept inclusion. The synonymy relation syn is construed as both way isa, cf. the declaration pancreas syn (pancreatic gland). Such propositions form the backbone of the ontology in our knowledge-based bio-models. Also partonomic propositions like betacell part-of (endocrine pancreas) are included in the ontology; cf. [8] for the various partonomic relations.

By contrast, a proposition like betacell secrete insulin is understood as an observational fact, an assertion, and therefore does not belong to the ontology proper. The concept of betacell would then be expected to be defined in some other way, which may or may not be part of the logical bio-model. However, the stated assertion might be replaced by the definitional proposition (cell that secrete insulin) syn betacell at the discretion of the domain expert. This proposition posits that all cells that secrete insulin (whatever their location), are to be called betacells.

3 Bio-models as Semantic Graphs

In our framework, the natural logic propositions constituting a bio-knowledge base are parallelled by an alternative representation in the form of directed graphs as commonly used in bio-models [9–11]. The graphs come about by decomposing compound and relational concept terms into their constituents in the form of triples [6]. These triples are re-conceived of as labeled directed edges between nodes. Every concept is associated with one node and *vice versa*.

This semantic graph representation facilitates computation of relevant associations between concepts, namely by computation of connecting paths in the graph. For example, the subject concept in the proposition (cell that secrete insulin) located-in pancreas corresponding to the natural language sentence *cell that secretes insulin is located in pancreas* is internally decomposed into the two triples

(cell-that-secrete-insulin) isa cell.
(cell-that-secrete-insulin) produce insulin.

where the added auxiliary concept (cell-that-secrete-insulin) is conceived of as an atomic name of a node defined by the two triples. An arc symbol as in '$\lhd$' is inserted between the defining edges in the graph rendition to express that they form the definition of the concept, in casu (cell-that-secrete-insulin).

The given proposition, which is epistemically in observational mode, then becomes represented by the triple (cell-that-secrete-insulin) located-in pancreas. So in this way a distinction is made between definitional and assertive (observational) propositions. This ensures that the original propositions, whatever their complexity, can be reconstructed modulo paraphrasation from the semantic graph as indicated with the double-headed arrow in figure 3. A graph representing the proposition and the decomposed subject term is shown in figure 1.

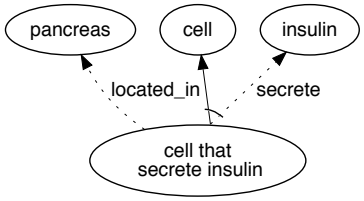

Fig. 1: The graph corresponding to the natural language sentence *cell that secretes insulin is located in pancreas.*

4 Fragments of a Bio-model: A Case Study

To exemplify the approach, below we develop fragments of logical knowledge base representation based on excerpts from Wikipedia articles on endocrine glands and insulin. The fragments of concern are stated as propositions in an extended, relaxed form of NaturaLog, cf. [3].

Some propositions introduce sub-concepts by agglutination rather than by the use of separate words, calling for manual treatment. Conversely, some would-be compounds like *islet of Langerhans* and *Graves' disease* should not be decomposed, but should be kept as atomic class names. From the source [12] we consider the following

> ***Endocrine glands are glands*** *of the endocrine system* ***that secrete hormones*** *directly into the blood rather than through a duct. The major* ***glands of the endocrine system include*** *the pineal gland, pituitary gland,* ***pancreas****, ovaries, testes,* ***thyroid gland****,* ***parathyroid gland****,* ***hypothalamus*** *and adrenal glands.*

leading to the following triples

> (endocrine gland) isa gland.
> (endocrine gland) isa (gland that secrete hormone).
> pancreas isa (endocrine gland).
> hypothalamus isa (endocrine gland).
> (thyroid gland) isa (endocrine gland).
> (parathyroid gland) isa (endocrine gland).

where a few of the obvious triples corresponding to propositions about location of the secretion and further specialisations of (endocrine gland) are omitted. Corresponding to these propositions we derive the semantic graph shown in figure 2. In addition, also from [12], we consider:

> *The* ***pancreas****, located in the abdomen close to the stomach,* ***is both an exocrine and an endocrine gland****.* ***Calcitonin****,* ***produced by*** *the* ***parafollicular cells of the thyroid gland in response to rising blood calcium levels****, depresses blood calcium levels by inhibiting bone matrix resorption and enhancing calcium deposit in bone.*
> *The* ***parathyroid glands****, located on the dorsal aspect of the thyroid gland,* ***secrete parathyroid hormone****, which* ***causes an increase in blood calcium levels****.*

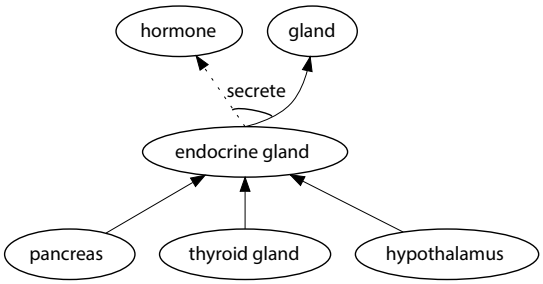

Fig. 2: Semantic graph focussing on endocrine gland.

from which we derive the following propositions (again omitting some to limit the extent of the example):

pancreas isa (endocrine gland).
pancreas isa (exocrine gland).
calcitonin produced-by (parafollicular cell in the thyroid gland).
(parafollicular cell in the thyroid gland) located-in (thyroid gland).
(parafollicular cell in the thyroid gland) isa (parafollicular cell).
(parafollicular cell) isa (cell).
(rising calcium level in blood) cause (production of calcitonin).
(production of calcitonin) produce calcitonin.
(rising calcium level in blood) located-in blood.
(parathyroid glands) secrete (parathyroid hormone).
(parathyroid hormone) cause (rising calcium level in blood).

All derived propositions above, including those shown in figure 2, are situated in the semantic graph shown in figure 4.

5 A prototype system

As indicated above, pathway query answers are provided by first extracting propositions from relevant texts as contributions to the semantic graph. The extracted propositions are combined with knowledge from supplementary sources into a semantic graph with unified nodes. Finally, pathways are computed in a separate module based on the semantic graph. We briefly describe these tasks below.

5.1 Extracting propositions from text

The problem of deriving NATURALOG propositions from a natural language text remains an open issue. Our main idea is to analyse the text seen from an extended version of NATURALOG (see [3]). This extension is purely syntactical so that it captures more expression forms in the text. However, it is at present semantically

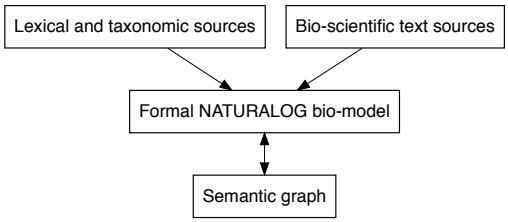

Fig. 3: Building the semantic graph.

conservative in the sense that propositions in the extended NaturaLog can be decomposed into the simple NaturaLog applied here. We intend to pursue this approach by stepwise extending also semantically NaturaLog to capture more meaning in the text. For instance, various forms of anaphora constructs fall outside NaturaLog, semantically. Also, for the moment, we consider only affirmative propositions, although in bio-texts one comes across negations for instance in the form of exceptions.

Furthermore, we envisage that one sentence may give rise to multiple propositions e.g. due to linguistic conjunctions, appositions, and parenthetical relative clauses. As shown in figure 1, one proposition in general gives rise to multiple triples in the graph rendition by a decomposition introducing nodes for compound, auxiliary terms.

5.2 Building the semantic graph

Given a proposition extraction module, a semantic graph is built by processing a corpus and situating the extracted triples in the graph. The graph is incrementally expanded by results derived from supplementary texts. Apart from textual input, contributions to the knowledge base may also be in the form of common knowledge lexical ressources (such as WORDNET), domain specific structured vocabularies/thesauri (such as UMLS) as well as other medical and bioscience sources that include taxonomic knowledge. Common for these sources is that they provide what can be considered concepts and relations connecting these. Therefore, the transformation into NaturaLog triples can be done by simple means. These triples, however, connect only atomic concepts. Thus, the contributions from such resources can be considered "skeleton"-ontologies to be further expanded with new atomic and compound concepts extracted from textual sources. The resource-based skeleton ontology is thus expanded into a generative ontology that grows incrementally with concepts and triples derived from new text sources. A sketch of the ontology building is shown in figure 3.

In figure 4 an example semantic graph is shown. The graph includes the example propositions derived in section 4 and it includes the following additional atomic concept triples:

gland isa organ.
stomach isa organ.

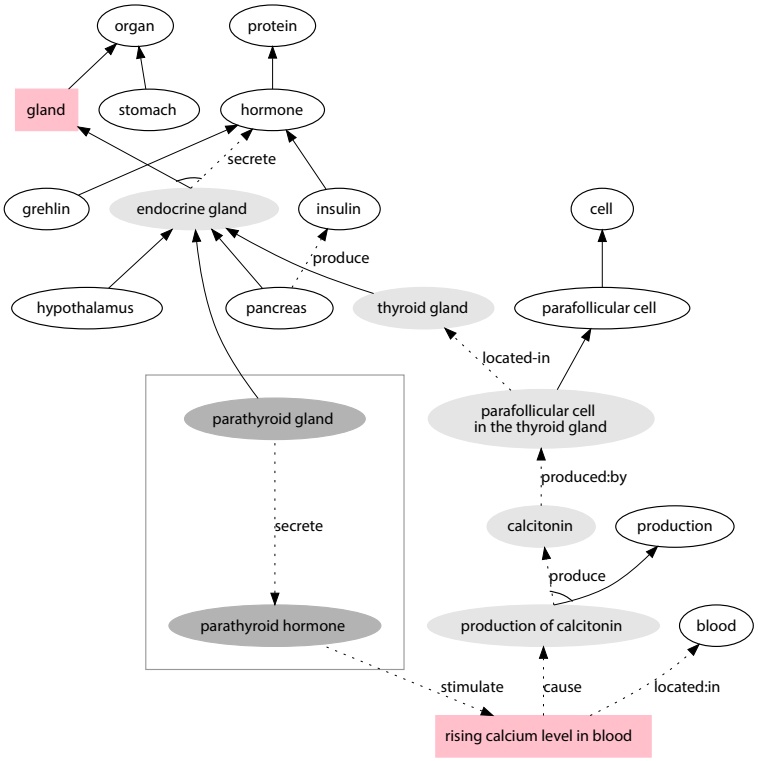

Fig. 4: A miniature ontology corresponding to a subset of the bio-model propositions listed in section 4. Two pathways connecting "rising calcium level in blood" and "gland" are shown.

grehlin isa hormone.
hormone isa protein.
insulin isa hormone.
pancreas produce insulin.

These may stem from results of other text sources or may be assumed to be part of a skeleton ontology that forms the basis for building the semantic graph.

5.3 Computational Query Answering and Pathfinding

Relationships derived by specialization of the subject and generalization of the object, know as inference by monotonicity, are identified by computational traversal of stated relationships. Concepts are connected in the semantic graph by pathways reflecting mathematical composition of the relations represented by edges in the logical bio-model, cf. [2]. It is our tenet that bio-pathways appear among the computed pathways between the given query concepts.

This computation process is supported by logical inference rules since inferred propositions may constitute shortcuts, as it were, in the graph view. For

instance, the transitivity of inclusion, isa, conceptually shortens the distance from a concept to a superior concept in the ontology via intermediate concepts. Similarly for partonomic, causative and effect relations. In [4], the path finding is explained more abstractly as application of appropriate logical comprehension principles supporting the relation composition.

A miniature ontology, corresponding to a subset of the bio-model propositions listed in section 4, is visualised in the graph in figure 4. In addition, two candidate answers to the query comprising two concepts

rising calcium level in blood $\sim$ gland?

are indicated.

An answer is provided as a pathway connecting the two concepts and the pathway can be seen as an explanation of how the two concepts are related. Consider the graph in figure 4 and assume that the darkgrey nodes are not yet inserted. The reading of the (answer corresponding to) lightgrey path can be

> *Rising calcium level in the blood causes production of calcitonin in the parafollicular cells in the thyroid gland, which is an endocrine gland, which is a gland.*

Suppose, at a later state, that new knowledge is being added to the base and that this include the darkgrey nodes in figure 4. The two query concepts are now connected by a new and shorter pathway corresponding to the following alternative answer.

> *Rising calcium level in the blood is stimulated by parathyroid hormone, which is secreted by the parathyroid gland, which is an endocrine gland, which is a gland*".

A pathway computation, being more than a pure inferential process, in our system is also the composition of relations guided by appropriate path computation. In our framework this computation is reduced algorithmically to search for weighted paths between concept nodes in the graph representation, utilizing standard heuristic algorithms in artificial intelligence. The intermediate propositional representations refer back to the source texts so that computed paths can be shown by highlighting excerpts in the texts.

6 Summary and Conclusion

We have described a system for querying and pathfinding in bio-models taking the form of logical knowledge bases derived from text sources. The applied logical language accommodates complex propositions, which can be queried by deductive means, and the supporting semantic graph form enables algorithmic pathfinding between concepts. A small scale prototype has been developed that translates complex propositions into a graph representation for pathfinding. This prototype is described in detail in [5]. Computational translation of text sources into the logical form is a challenging problem, which is approached in [3] by adopting enriched forms of natural logic as a specification language for biosystems.

References

1. Schultz, S. & Hahn, U.: Towards the ontological foundations of symbolic biological theories, *Artificial Intelligence in Medicine*, 2007, 39, 237-250.
2. Bittner, Th. & Donelly, M.: Logical properties of foundational relations in bio-ontologies: *Artificial Intelligence in Medicine*, 2007, 39, 197-216.
3. Andreasen, T., Bulskov, H., Fischer Nilsson, J., Jensen, P.A.: On the Relationship between a Computational Natural Logic and Natural Language In: ICAART The 8th International Conference on Agents and Artificial Intelligence 2016: 335-342
4. Andreasen, T., Bulskov, H., Fischer Nilsson, J., Jensen, P.A., Lassen, T.: Conceptual Pathway Querying of Natural Logic Knowledge Bases from Text Bases. In Proceedings of the 10th international conference on Flexible Query Answering Systems, Springer-Verlag, Berlin, Heidelberg (2013) 1-12
5. Andreasen, T., Bulskov, H., Fischer Nilsson, J., Jensen, P.A.: A System for Conceptual Pathway Finding and Deductive Querying. In Proceedings of the 11th international conference on Flexible Query Answering Systems, Springer-Verlag, Berlin, Heidelberg (2015)
6. Fischer Nilsson, J.: Diagrammatic Reasoning with Classes and Relationships, Moktefi, A. & Shin, S.-J. (eds.) *Visual Reasoning with Diagrams*, Studies in Universal Logic, Birkhäuser, Springer, 2013.
7. van Benthem, J.: Essays in Logical Semantics, Studies in Linguistics and Philosophy, Vol. 29, D. Reidel Publishing Company (1986)
8. Smith, B. & Rosse, C.: The Role of Foundational Relations in the Aligment of Biomedical Ontologies, MEDINFO 2004, M. Fieschi *et al.*, 2004.
9. Vechina, A. *et al.*: Representation of Semantic Networks of Biomedical Terms. In Proceedings of the International Work-Conference on Bioinformatics and Biomedical Engineering, Granada, Spain, 2013.
10. Miljkovic, D. *et al.*: Incremental revision of biological networks from texts. In Proceedings of the International Work-Conference on Bioinformatics and Biomedical Engineering, Granada, Spain, 2013.
11. Quesada-Martinéz, M. *et al.*: Analysis and Classification of Bio-ontologies by the Structure of their Labels. In Proceedings of the International Work-Conference on Bioinformatics and Biomedical Engineering, Granada, Spain, 2013.
12. Endocrine gland (2015), Retrieved March 6, 2016, from *Wikipedia* `http://en.wikipedia.org/wiki/Endocrine_gland` (2015)
13. Li, C., Liakata, M., and Rebholz-Schuhmann, D. (2013). Biological network extraction from scientific literature: state of the art and challenges. Briefings in Bioinformatics.
14. Kaewphan, S., Kreula, S., Van Landeghem, S., Van de Peer, Y., Jones, P. R., and Ginter, F. (2012). Integrating Large-Scale Text Mining and Co-Expression Networks: Targeting NADP(H) Metabolism in E. coli with Event Extraction. In Proceedings of the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2012), pages 8-15.
15. Hakala, K., Van Landeghem, S., Kaewphan, S., Salakoski, T., Van de Peer, Y., and Ginter, F. (2012). CyEVEX: Literature-scale network integration and visualization through Cytoscape. In Proceedings of SMBM'12, Zurich, Switzerland, pages 91-96.
16. Miwa, M., Ohta, T., Rak, R., Rowley, A., Kell, D. B., Pyysalo, S., and Ananiadou, S. (2013). A method for integrating and ranking the evidence for biochemical pathways by mining reactions from text. Bioinformatics 29(13):i44-i52.