# Towards Knowledge Discovery through Automatic Inference with Text Mining in Biology and Medicine

**Hee-Jin Lee and Jong C. Park**
Computer Science Division, KAIST
373-1 Guseong-dong, Yuseong-gu, Daejeon 305-701 Republic of Korea
{heejin,park}@nlp.kaist.ac.kr

## Abstract

Field experts in biology and medicine search the literature for state-of-the-art results and occasionally discover knowledge through manual inference on published causal relations. However, the results of such inference cannot be sufficiently accurate and/or complete, as the domain of published relations is rather huge. In this paper, we introduce an automatic inference system, BioDetective, which works on literature-mined qualitative causal information in biology and medicine. BioDetective provides proofs for such qualitative causal information, and predicts the existence of new causal information, if there is any. The system is tested with a case study, where literature-mined information about protein regulation is utilized to come up with new knowledge.

## 1 Introduction

Field experts in biology and medicine search the literature for state-of-the-art results and occasionally discover knowledge through manual inference on published causal relations. For example, a biomedical scientist who seeks new treatments for a disease may search the literature for information about biological or medical entities already known to be related to the particular disease, as well as about causal relations that are known to involve these entities. By inferring over the combined effect of such causal relations, she may be able to discover novel (and possibly indirect) causal relations between the disease and some molecules and/or biological conditions. She may also use such novel causal information towards finding effective drugs for the disease. Such an approach to knowledge discovery through manual inference on literature-mined information would be a good way to reduce the number of repeated experiments that are based purely on intuition, which may turn out to be not only time-consuming but also literally quite expensive.

However, the fraction of information that can be manually examined this way is much limited. For one, the experts may not be able to locate the connecting information that would have been easily identified if the available body of knowledge were larger. Even when a larger body of knowledge is taken into account, manual inference is susceptible to mistakes due to the complexity of the involved inference. Automated inference on a dataset of literature-mined information will certainly help the field experts to cover more information with fewer mistakes.

In this paper, we introduce an automatic inference system, BioDetective, for literature-mined qualitative causal information in biology and medicine. Given a collection of causal information and other related literature-mined information as the input dataset, BioDetective can check if the input dataset supports a new, hypothetical causal relation between known biological (or medical) entities. We believe that this helps field experts to discover new knowledge from literature-mined information. In order to assess the performance, we tested the system with a case study, where literature-mined information about protein regulation is employed.

We review other inference systems in Section 2, introduce BioDetective in Section 3, describe our case study in Section 4, and show concluding remarks in Section 5.

## 2 Related Work

| Notation | Types | | | Description | Flow of effects |
|---|---|---|---|---|---|
| | x | y | z | | |
| z | | | SE | Biological Process | Ⓩ |
| z | | | SEM | Molecule | Ⓩ |
| z | | | SE | External control or disease | Ⓩ |
| z → y | | M | SEM | Modification to molecule | Ⓩ ← Ⓨ |
| x ← z → y | M | M | SEM | Binding | Ⓧ → Ⓩ ← Ⓨ |
| x — z → y | S | E | E | Inhibition | Ⓧ → Ⓩ → Ⓨ |
| x — z ▷ y | S | E | E | Induction | Ⓧ → Ⓩ → Ⓨ |
| x — z ▷\| y | S | E | | Necessary condition | Ⓧ → Ⓩ → Ⓨ |
| x — z → ø | S | S | E | Degradation | Ⓧ ← Ⓩ |

Table 1. Symbols used in DPL. S is for states, E for events, and M for molecules.

BioDetective handles high level concepts together with molecular level concepts, as the information in the literature is often at a level higher than the molecular level. The system can also accommodate new kinds of concepts, unknown to the system beforehand. These two characteristics, unavailable from current systems, of which some are reviewed below, facilitate the system to produce new information from literature-mined information, by enabling the system to connect information which would be considered otherwise unrelated.

BioSigNet-RR is a system for representing and reasoning about signaling networks (Baral et al., 2004), and can deal with four kinds of queries on the cellular signaling network, but does not seem to be easily adaptable to other sub-domains of biology. BIOCHAM is a software tool for modeling biochemical systems (Calzone et al., 2006), and can conduct analysis and simulation of biological models, but requires data to be only at the molecular level. Pathway Logic is an approach to modeling biological entities/processes based on rewriting logic (Eker et al., 2005), for the analysis of models of signal transduction networks, but does not appear applicable to literature-mined information with higher level concepts.

## 3 BioDetective

Given a database of biological causal information and a query describing a causal relation, BioDetective checks if the input dataset supports the causal relation[1]. The structure of BioDetective is shown schematically in Figure 1.

The input database should contain a dataset that forms a causal network, to be defined in Section 3.1. The information in the input database is used by the model generator to generate a model

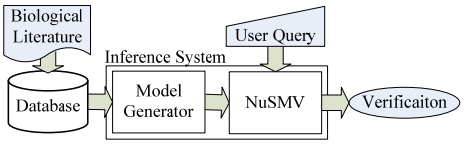


Figure 1. Structure of BioDetective

of a concurrent system, which follows rules in Section 3.2. The generated model is provided to NuSMV (Cimatti et al., 2002), an open source model checker. A causal relation stated as a temporal query as explained in Section 3.3 is provided to NuSMV for verification.

### 3.1 Causal network as input

To describe the datasets that can be used as input to BioDetective, we define Diagrammatic Pathway Language (DPL) (cf. Park and Park, 2005).

DPL is a set of pathways, where a pathway is defined as a set of connected symbols, which are any of the symbols of 9 types shown in Table 1, where positions marked by x and y are instantiated by other connected symbols. DPL follows the style as proposed by Kohn and others (2006).

A collection of biological or medical information forms a causal network when each piece of such information is represented with a corresponding symbol in DPL. We assume that the body of information comprising the input dataset forms a causal network.

Notice that the information in a causal network includes higher level concepts such as causal relations, and non-causal concepts such as diseases and biological processes.

### 3.2 Rules for concurrent systems

The concurrent system of a causal network generated by the model generator consists of biological or medical entities represented each with a symbol in the pathway of the causal network. Each entity of the concurrent system can either be present or absent. The status of entities may change simultaneously over time according to

[1] The system is based on the abstract description of a qualitative formalization framework by Park and Park (2005), implemented here with extensions for automated execution.

| Rule name | Description |
|---|---|
| Environment Assumption | 1. When a non-external molecule A is not the target of any induction or inhibition, the molecule is considered initially present.<br>2. When a disease or external control or a molecule is the target of any induction or inhibition, it is considered initially not present.<br>3. Otherwise, the status of the biological entity is considered not initially determined. |
| Implicit Necessary Condition | The presence of participants of a biological entity is a necessary condition for the biological entity. |
| Dynamic Inference | - Biological entity X will be present if<br>1. for some A that induces X, A is present, and<br>2. for all B that inhibits X, B is absent, and<br>3. for all necessary conditions C for X, C is present.<br>- Biological entity X will be absent, otherwise. |
| Inertia | Once a biological entity becomes present by the Dynamic Inference rule, it remains present unless it is interfered. |

Table 2. Rules for concurrent systems.

the causal relations involving the entities. Rules for such status changes are summarized in Table 2. The status changes of all the entities in a concurrent system reflect the combined effect of all the causal relations in the causal network.

Note that the rules are not specific to the kind of entities involved in the chain of causal relations, and that the inference system can easily accommodate new types of biological entities.[2]

### 3.3 Causal relations as queries for NuSMV

We use NuSMV to compute the temporal properties of a concurrent system to verify causal relations of interest. A causal relation between two biological or medical entities is stated as temporal properties in Linear Temporal Logic (LTL), using two LTL operators 'in the future (*F*)' and 'globally (*G*)'. Given a formula in LTL as a query, NuSMV returns *true* if the concurrent system of the input causal network has the queried temporal property, which means that the queried causal relation is supported by the input dataset. Inducing and inhibiting relations between entities A and X are stated respectively as follows.

- Induction of *X* by *A*: *A -> F G X*
- Inhibition of *X* by *A*: *A -> F G !X*

If the causal relation is not shown explicitly with a symbol in the pathway of the causal network, the relation is considered indirect. An indirect causal relation verified by BioDetective would work as a novel piece of information, obtained by connecting pieces of known information.

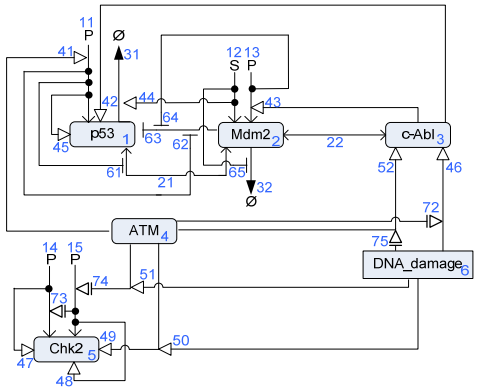


Figure 2. The causal network constructed as in Section 4. Every entity is marked with its ID number.

## 4 Using BioDetective: A case study

In this section, we demonstrate possible uses of BioDetective with a case study, where the system takes literature-mined information as input and produces new knowledge, if there is any.

**Construction of a causal network:** We constructed a causal network semi-automatically so that it can be used as input for our case study, as shown in Figure 2. Using BioIE (Kim and Park, 2004), an information extraction system specialized for biology and medicine, we extracted 109 descriptions of interactions and causal relations between ATM, Mdm2, Chk2, c-Abl, p300 and p53, from MEDLINE. We then manually examined the extracted pieces of information to construct a causal network and manually stored the network in an SQLite database.[3]

**Phase 1 – Resolving multiple representations:** There are cases where a causal relation is represented explicitly in a causal network, but is also represented by paths of other causal relations and interactions in the same network. These cases possibly result from the mixed nature of

[2] We believe that the 9 types of biological entities currently handled by the system form a complete set of types, but new types such as 'phenotypes' may still be introduced if needed. Existing types such as 'induction' may also be split into lower level types, such as 'induction by transcription'. However, the rules themselves remain unchanged.

[3] The causal network contains 34 biological entities; We did not use all the extracted information.

natural language descriptions, each describing the same event with a different level of detail.

We utilized BioDetective to detect these cases. Given causal relation I, where A is a direct cause for B to change, we collected all the causal relations and interactions that transfer the status of A to B, to construct a subnetwork. This is done by using the information in the last column of Table 1. We then used BioDetective to see if the subnetwork supports I. If BioDetective returns *true*, the subnetwork is interpreted as representing the same event as I, but at a level more detailed than the one suitable for representing I.

We applied the procedure above to the input causal network as shown in Figure 3. We found five explicit causal relations having a subnetwork of the same effect. One of them is the inducing relation 42, 42 being the ID number in Figure 3, where the corresponding subnetwork consists of biological entities 1, 63, 64, 13, 43, 2, and 3. This multiplicity was evidenced by the following sentence found manually from the literature.

- Phosphorylation of Mdm2 by c-Abl impairs the inhibition of p53 by Mdm2, hence defining a novel mechanism by which c-Abl activates p53. [PMID: 12110584]

We removed all the five explicit causal relations to use the modified network in phase 2.

**Phase 2 – Finding sufficient conditions for events to happen:** If we consider a causal network of literature-mined information as a qualitative model of a biological system, and use a closed world assumption, we can find a sufficient condition for a biological or medical event to happen, using BioDetective.

The sufficient condition for a biological or medical entity X to be present can be found by searching for the initial configuration A of the input causal network that supports the query 'A -> F G X'. For this purpose, the causal network in Figure 3, cleaned up at the first phase of the case study, was used. One of the sufficient conditions found by the system is shown below.

- Absence of Mdm2 at the initial time is a sufficient condition for p53 to be present.

We plan to improve the system performance further by selecting a subset of the configurations of the initial status of the input network.

## 5 Concluding Remarks

We introduced BioDetective, an automatic inference system that deals with qualitative causal information in biology and medicine. The system is suitable for producing new information by meaningfully connecting existing pieces of information in the literature.

The system is utilized in a case study, where literature-mined information is collected and processed to obtain new knowledge. The case study showed the possibility that the system is applicable for various tasks for integration and utilization of the literature-mined information.

## Acknowledgments

We thank Robert Kueffner and Hodong Lee for valuable comments on earlier versions of this paper. This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF-2007-313-D00738) and by the research funding on an opportunistic project by Microsoft Research Asia.

## References

Chitta Baral et al. 2004. A knowledge based approach for representing and reasoning about signaling networks, *Bioinformatics*, 20:i15-i22.

Laurence Calzone et al. 2006. BIOCHAM: an environment for modeling biological systems and formalizing experimental knowledge, *Bioinformatics*, 22(14):1805-1807.

Alessandro Cimatti et al. 2002. NuSMV 2: An open-source tool for symbolic model checking, *In Proceedings of CAV 2002*, 27-31.

Steven Eker et al. 2005. Pathway Logic: executable models of biological networks, *Electronic Notes in Theoretical Computer Science*, 71: 144-161.

Jung-jae Kim and Jong C. Park. 2004. BioIE: retargettable information extraction and ontological annotation of biological pathways from the literature, *Journal of Bioinformatics and Computational Biology (JBCB)*, 2(3):551-568.

Kurt W. Kohn et al. 2006. Molecular interaction maps of bioregulatory networks: A general rubric for systems biology, *Molecular Biology of the Cell*, 17:1-13.

Il Park and Jong C. Park. 2005. Modeling causality in biological pathways for logical identification of drug targets, *The 2005 International Joint Conference of InCoB, AASBi and KSBI (Bioinfo 2005)*, 373-378.