

# Modeling Causality in Biological Pathways for Logical Identification of Drug Targets

Il Park Jong C. Park

Computer Science Division, KAIST, Daejeon, South Korea

Email: memming@nlp.kaist.ac.kr, park@nlp.kaist.ac.kr

**ABSTRACT:** The diagrammatic language for pathways is widely used for representing systems knowledge as a network of causal relations. Biologists infer and hypothesize with pathways to design experiments and verify models, and to identify potential drug targets. Although there have been many approaches to formalize pathways to simulate a system, reasoning with incomplete and high level knowledge has not been possible. We present a qualitative formalization of a pathway language with incomplete causal descriptions and its translation into propositional temporal logic to automate the reasoning process. Such automation accelerates the identification of drug targets in pathways.

## 1 INTRODUCTION

The diagrammatic language of biological pathways represents the systems knowledge of a biological process, or a collection of related causal relations. Biologists use pathways to represent not only their current understandings but also hypotheses of the system. Furthermore, inference over the pathways gives rise to new pieces of knowledge about and insight into the biological process.

Despite efforts to standardize the diagrammatic language for pathways [1, 2, 3, 4], the resulting proposals are not yet widely used in the biological field. Most of these efforts aim to mathematically model the pathways, and focus on the complete description and simulation of the system. Nevertheless biologists still use an informal representation for pathway, called in this paper *informal pathways*. It is because the biological knowledge is still incomplete and insufficient for full simulation [5], and the causality of other levels cannot be naturally incorporated into these standardized pathway languages. Nonetheless, biologists are still able to understand, communicate, hypothesize and reason with informal pathways as frequently observed from the literature.

The mixture of complete mechanical knowledge and qualitative causal knowledge is both represented in the informal pathways. The use of pathways in modeling the current knowledge from experiments and hypothesizing for the next experiment is common in biology [6]. Causality is a critical component to the iterative cycle of experiment and modeling. In addition, informal pathways are used to model the cause of a disease and to predict plausible drug targets. This part of drug discovery process heavily relies on the causal relations for a com-

pact and flexible reasoning rather than on the complete mechanical knowledge [7].

In this paper, we propose an expressive pathway language that combines different levels of causal knowledge, and a logical inference system that mimics the inferences on the pathway as performed by the biologists.

## 2 CAUSAL PATHWAY

Instead of the complete molecular description, the informal pathway contains indirect causal relations, namely, induction and inhibition. Induction of an event is a general positive regulation and its inhibition is a general negative regulation. Although some molecules or interactions may be underlying the induction or inhibition, this partial information is still useful for qualitative reasoning. The pathway language we propose in this paper is a hybrid of molecular level interactions and high level concepts to represent the informal pathways as used by the biologists.




Notation	Types			Description
	X	Y	Z	
			SE	Biological process
			SEM	Molecule (non-external)
			SE	External control or disease
$x \xrightarrow{z} y$	O	M	SE	Modification to molecule
$x \xleftarrow{z} y$	M	M	SEM	Binding
$x \xrightarrow{z} \triangleright y$	S	E	E	Induction
$x \xrightarrow{z} \dashv y$	S	E	E	Inhibition
$x \xrightarrow{z} \dashv\triangleright y$	S	E		Necessary condition
$x \xrightarrow{z} \blacktriangleright y$	S	S	E	Conversion

Table 1: Notations and corresponding type constraint used in the pathway. S is for states, E for events, O for modifications, and M for molecules.

The notation used in this paper is motivated by the Kohn interaction map [1, 5]. However, there are some

unification of the symbols for a simpler description and clarifications for unambiguous semantics. The basic symbols are defined in Table 1. Each node and edge is assigned a set of types, and the edges in a pathway are connected with respect to these types. For an edge to be connected, each end node should be a fragment of a well-formed pathway, so that there would be no infinitely recursive structure.<sup>1</sup>

Unlike pathway languages that concentrate only on molecular interactions, our approach also introduces non-causal higher level concepts such as diseases, experimental conditions, and biological processes. These higher level concepts often fill the gap among the molecular states that would be otherwise considered unrelated at the molecular level.

In Figure 1, we reconstructed the pathway of [8] using our notation that explains the experimental results. The pathway explains how estrogen receptor negative (ER-) breast cancer cells show an enhanced proliferation and how each experiment disturbs the process. A sequence of induction and inhibition from EGFR to cell proliferation and the possible external control of the sequence by several drugs demonstrate the causal chaining that was modeled by the biologists.

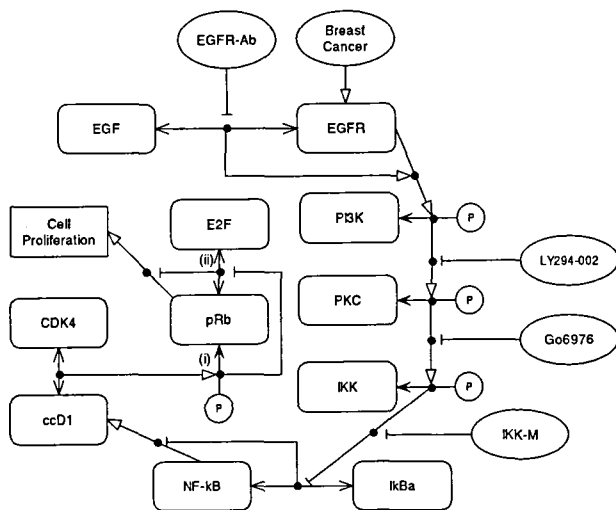


Figure 1: A pathway for the mechanism of ER(-) breast cancer. The roman numbers are for future reference.

### 3 ANALYSIS OF CAUSAL INFERENCE

The chain of causal inference from EGF and breast cancer to cell proliferation, which implies that the pathway is abused by the cancer, can be obtained by using some informal rules of inference over the pathway. We assume that each state or event can be either *present* or *absent*; a present state or event means that the biological system

<sup>1</sup>Both ends of an edge should be fully instantiated without the edge. For example, connections such as  $\rightarrow$  are meaningless in a pathway diagram.

contains the corresponding physical state or event independent of other states or events, and the converse for an absent state or event. For example, if we say that the state for phosphorylated PIK3 is present, the biological system currently contains phosphorylated PIK3, and it is independent of the presence of other states such as a state for phosphorylated EGFR which might be absent at the moment.

To represent the initial condition, each symbol for a state has a predefined semantics. Initially, the rounded-boxed molecules except for ccD1 in Figure 1 are assumed to be present since their presence is independently motivated. And the presence of the ovals is undetermined, since they can be controlled externally. This can be encoded into the following rule.

#### Rule 1 (Environment Assumption)

1. When a non-external molecule *A* is not a target of any induction or inhibition, the molecule is initially present.
2. When a disease or external control or a molecule is a target of any induction or inhibition, it is initially not present.
3. Otherwise, the presence of the states is not initially determined. (Undetermined states)

For a clear representation, obvious necessary conditions for events, that is the presence of the participants that is implied in the diagram, are omitted. For example, the binding of EGF and EGFR would require the presence of both EGF and EGFR, and the induction of the phosphorylation of PI3K by EGFR requires the presence of EGFR.

**Rule 2 (Implicit Necessary Condition)** *The presence of participants of a state or event is a necessary condition for the state or event.*

Intuitively, direct inductions and inhibitions can be chained to entail indirect conclusions. For example, the signaling path along the kinases, EGFR, PI3K, PKC and IKK, is connected via consecutive inductions, and the first induction entails the last induction, so that the signal is transmitted through phosphorylated states. The next rule states the generalization of the idea.

#### Rule 3 (Chaining of Induction and Inhibition)

*Assuming *X* and *Y* have no other induction, inhibition or necessary condition for *X* or *Y*,*

- if induction of *X* by *Y* is induced by *Z*, then *X* is indirectly induced by *Z* in the presence of *Y*;
- if inhibition of *X* by *Y* is induced by *Z*, then *X* is indirectly inhibited by *Z* in the presence of *Y*;
- if induction of *X* by *Y* is inhibited by *Z*, then *X* is indirectly inhibited by *Z* in the presence of *Y*; and
- if inhibition of *X* by *Y* is inhibited by *Z*, then *X* is indirectly induced by *Z* in the presence of *Y*.

However, in a general network of induction, inhibition, and necessary condition, the multiplicity of the connections prevents the direct application of Rule 3. The general dynamic property concerning the causal relation can be given by the following rule. Rule 3 is a special case of Rule 4.

#### Rule 4 (Dynamics Inference)

- State or event  $Y$  will be present if
  1. for some  $X$  that induces  $Y$ ,  $X$  is present,
  2. for all  $Z$  that inhibits  $Y$ ,  $Z$  is absent, and
  3. for all necessary condition  $P$  for  $Y$ ,  $P$  is present, and
- State or event  $Y$  will be absent if
  1. for at least one  $X$  that inhibits  $Y$ ,  $X$  is present, and
  2. for at least one  $X$  that is a necessary condition for  $Y$ ,  $X$  is not present.

For example, the state for phosphorylated pRb will be present if CDK4 and ccD1 binds, and the state for binding of E2F and pRb will be absent (disassociated), if pRb is phosphorylated, marked as (i) and (ii), respectively, in Figure 1.

While some of the states change, there are also static states. Unless there is any reason to change the presence of the state, it remains unchanged.

**Rule 5 (Inertia)** Once a state became present by Rule 4, it remains present unless it is interfered.

Using these informal rules, it is possible to infer that the breast cancer cells will proliferate and that the inhibitory drugs or experiments will block this effect, and therefore, to discover that the molecules involved in the pathway are potential drug targets used in the experiment and that the drugs could be hypothesized as a cure for the cancer.

## 4 TRANSLATION INTO PROPOSITIONAL TEMPORAL LOGIC

The informal rules allow informal inferences; however, translation into a well founded logic provides a concrete semantics and automated theorem proving. Moreover, the inference over a pathway is non-monotonic, in the sense that newly introduced edges can make a valid inference invalid, which generally makes the inference system complex. In order to deal with non-monotonicity, our approach is to assume a closed world, and to translate expressions in the pathway language into ones in a monotonic logic, propositional temporal logic (PTL).

Modeling causality with logic has been a difficult problem [9], and there have been many modal logic approaches such as counterfactuals [10], dynamic logic [11], and action logic [12]. Since the causality in the biological domain is more restricted and precise than the general

causality, it may be modeled through a simpler framework, especially temporal logic which is widely used to represent the temporal relationships.

The basic syntax of PTL with which we will be translating pathways is briefly defined below with propositional symbols and formulas. A propositional symbol will denote a state, and a formula represents an assertion about the temporal and logical relation among the states and events. Three modal operators, next time ( $\bigcirc$ ), sometime in the future ( $\diamond$ ) and always in the future ( $\square$ ), are used.

**Definition 1 (Proposition Symbols)** Let  $M$  be the set of molecules,  $EC$  the set of external controls, and  $BP$  the set of biological processes of a pathway. The set of propositional symbols for the pathway,  $P$ , is defined as follows.

$$P = MU\{A_P | A \in M\} \cup \{BIND_{A,B} | A, B \in M\} \cup EC \cup BP$$

$A_P$  denotes a phosphorylated species of a molecule  $A$ , and  $BIND_{A,B}$  denotes a bound species of two molecules  $A$  and  $B$ . Phosphorylation on multiple sites and molecule complexes with more than two molecules are not considered, since they do not significantly change the reasoning and it is easy to extend the symbols to cover them. However, our representation will suffer from the combinatorial increase of states when these notions are included as in other representations.

**Definition 2 (Formulas)**  $F \in P$  is an atomic formula of PTL, and if  $F_1$  and  $F_2$  are formulas of PTL,  $\bigcirc F_1, \diamond F_1, \square F_1, F_1 \wedge F_2, F_1 \vee F_2, \neg F_1, F_1 \supset F_2$  are also formulas of PTL.

The translation function maps the pathway fragments to PTL formulas. Since a pathway fragment has a non-recursive finite structure, the translation function generates a finite formula. The process can be divided into two recursive functions; atomic translation which maps a state or conversion to a formula, and causal translation which maps a state or event to a formula.

#### Definition 3 (Atomic Translation Function)

The atomic translation function  $\bar{(\cdot)}$  from a state or conversion of a pathway to a PTL formula is defined as the following table.

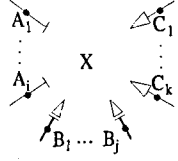
$\iota$	$\bar{\iota}$
$\boxed{X} \quad \boxed{x} \quad \bigcirc x$	$X$
$P \xrightarrow{\bullet} Y$	$Y_P$
$X \xleftrightarrow{\bullet} Y$	$BIND_{\bar{X}, \bar{Y}}$
$X \xrightarrow{\bullet} Y$	$\bar{X} \supset \bigcirc(\neg \bar{X} \wedge \bar{Y})$

In an atomic translation, a complex state is resolved to a propositional symbol and a conversion is resolved to a dynamic property. In a causal translation, the causal relations among the nodes are resolved, using a conjunction of four assertions corresponding to Rules 4 and 5. Rule 2 is assumed to be contained in the pathway before the translation.

#### Definition 4 (Causal Translation Function)

The causal translation function  $\widehat{(\cdot)}$  from a state or event  $X$  of a pathway to a PTL formula is defined as follows.

1. If  $X$  is a state or conversion, and no induction, inhibition, or necessary condition is connected to  $X$ ,  $\widehat{X} = \overline{X}$ ,
2. If  $X$  is an induction, inhibition, or necessary condition from  $Y$ , and no induction, inhibition, or necessary condition is connected to  $X$ ,  $\widehat{X} = \overline{Y}$ ,
3. Otherwise,



$$\begin{aligned} \widehat{X} &= (\neg(\bigvee_i \widehat{A}_i) \wedge (\bigwedge_j \widehat{B}_j) \wedge (\bigvee_k \widehat{C}_k)) \supset \diamond \overline{X} \wedge \\ &\quad ((\bigvee_i \widehat{A}_i) \vee (\neg \bigwedge_j \widehat{B}_j)) \supset \diamond \neg \overline{X} \wedge \\ (\neg \overline{X} \wedge \neg(\neg(\bigvee_i \widehat{A}_i) \wedge (\bigwedge_j \widehat{B}_j) \wedge (\bigvee_k \widehat{C}_k))) &\supset \bigcirc \neg \overline{X} \wedge \\ (\overline{X} \wedge \neg((\bigvee_i \widehat{A}_i) \vee (\neg \bigwedge_j \widehat{B}_j))) &\supset \bigcirc \overline{X} \end{aligned}$$

The pathway is translated into a formula containing all the static rules and dynamic rules of the pathway via the causal translation function. An explanation is forthcoming in 5.2 example. The initial conditions from Rule 1 are stated in the initial condition clause and the other rules are stated as a conjunction of translated states.

**Definition 5 (Initial Condition Clause)** Let  $\{TI_i\}$  and  $\{FI_j\}$  be the set of propositional symbols corresponding to Rule 1.1 and 1.2, respectively, given a pathway. The initial condition clause  $ICC$  is defined as follows.

$$ICC = \bigwedge_i TI_i \wedge \bigwedge_j \neg FI_j$$

**Definition 6 (Pathway Formula)** The pathway formula,  $PATHWAY$ , for a pathway is defined by the following formula,

$$PATHWAY = ICC \wedge \bigwedge_i \widehat{X}_i$$

where  $X_i$  is an undetermined state or conversion of the pathway.

Properties in question about the pathway can be examined by checking for the validity of a corresponding formula. Since the satisfiability problem of PTL is PSPACE-complete [13], and the validity of a given formula can be done by checking for the satisfiability of

the negation of the formula, they have the same complexity. Although most of the tautologies are vacuous without any biological meaning, the formulas of the form  $PATHWAY \supset X$  contain the properties represented by  $X$ .

## 5 RESULTS

### 5.1 Biological background

Recently, the inhibitor of PARP-1, which is involved in single strand break repair (SSBR), a DNA repair mechanism, is hypothesized and examined as a cure for breast cancer [14, 15]. Since only the cancer cells express the impaired BRCA1 gene, which is involved in double strand break repair (DSBR) via homologous recombination, the cells are unable to repair the DSBs. When the PARP1 inhibitor is treated, the SSB repair system malfunctions and more DSBs take place. Consequently, the treatment of the PARP1 inhibitor causes more DSBs to occur in cancer cells compared to normal cells, and the DSBs result in apoptosis selectively on cancer cells.

This sequence of reasoning process is interesting in that it involves non-trivial condition resolution. This pattern of reasoning could not easily done by other frameworks, and is also difficult to discover by the biologists when the size of the involved pathway is huge. The pathway that models the reasoning is shown in Figure 2. The details of the pathway unrelated to the reasoning we are focusing on are omitted for space reasons.

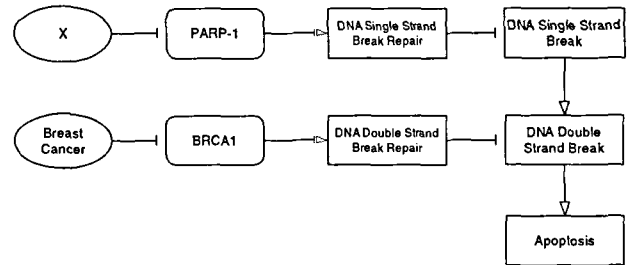


Figure 2: A high level DNA repair pathway related to breast cancer. The drug used as the inhibitor of PARP-1 is denoted as an external control  $X$ .

### 5.2 Translation and Inference

The translation of the pathway is straightforward. First, since there are no states corresponding to initial condition clause, it is empty and vacuously true. The remaining states  $\{PARP1, SSBR, SSB, BRCA1, DSBR, DSB, Apoptosis\}$  are then translated into  $PATHWAY$ . For example,  $DSB$  is translated into the following formula.

$$\begin{aligned}
& (SSB \wedge \neg DSB \supset \diamond DSB) \\
& \wedge (DSB \supset \diamond \neg DSB) \\
& \wedge (\neg DSB \wedge \neg (SSB \wedge \neg DSB) \supset \bigcirc \neg DSB) \\
& \wedge (DSB \wedge \neg DSB \supset \bigcirc DSB)
\end{aligned}$$

The partial translation above asserts that (a) the occurrence of single strand breaks to DNA without double strand break repairing would eventually give rise to double strand breaks, (b) the double strand break repair will prevent double strand breaks, (c) the absence of double strand breaks will continue if single strand breaks are absent or double strand break repairing is at work, and (d) the presence of double strand breaks will continue if there is a problem with double strand break repairing.

$$\begin{aligned}
& \text{PATHWAY} \supset \\
& (X \wedge \text{BreastCancer} \supset \diamond \text{Apoptosis}) \quad (1)
\end{aligned}$$

$$\begin{aligned}
& \text{PATHWAY} \supset \\
& (X \wedge \neg \text{BreastCancer} \supset \square \neg \text{Apoptosis}) \quad (2)
\end{aligned}$$

From the fully translated *PATHWAY*, we can infer the facts (1) and (2). (1) asserts that treating *X* to a breast cancer cell would eventually lead to apoptosis, and (2) asserts that treating *X* to a normal cell without breast cancer would not eventually lead to apoptosis. Thus we can conclude that *X* can selectively induce apoptosis on breast cancer cells, as predicted. The framework can support the identification of a drug target with a given pathway either by a fully automatic or semi-automatic hypothesis generation and verification.

## 6 CONCLUSION

We proposed a framework that mimics the reasoning of biologists on their pathway. Causal knowledge of mixed levels is represented in an unambiguous pathway language, and is translated into PTL for automated reasoning. The pathway language can naturally represent the incomplete causal knowledge and include high level concepts.

The primary focus of examples in this paper is on drug discovery. The reasoning process of the identification of drug targets and potential drugs is reconstructed and verified. The framework can also be applied to an automated experiment design system such as [16] for hypothesis generation and verification.

Pathway databases such as KEGG [17] or REACTOME [18] contain a lot of useful data. However, since they do not contain most of the high level causalities that our system can utilize, it is necessary to construct pathways from other sources. Since the pieces of causal knowledge are experimental results, they are reported in the literature as diagrams and in natural language. Using information extraction systems such as BioIE [19] to construct causal pathways and automated knowledge discovery is one of our next goals.

## ACKNOWLEDGEMENT

We would like to thank Ho-Joon Lee and Woosuk Park for valuable comments. This work was supported by grant No. R01-2005-000-10824-0 from the Basic Research Program of the Korea Science & Engineering Foundation.

## REFERENCES

- [1] Kurt W. Kohn. Molecular interaction map of the mammalian cell cycle control and DNA repair systems. *Molecular Biology of the Cell*, 10(8):2703–34, 1999.
- [2] Hiroaki Kitano. A graphical notation for biochemical networks. *BIOSILICO*, 1(5):169–176, 2003.
- [3] Michael Hucka, Andrew Finney, Herbert Sauro, and Hamid Bolouri. Systems biology markup language (SBML) level 1. 2003.
- [4] BioPAX workgroup. BioPAX biological pathways exchange language level 1, version 1.2 documentation. 2004.
- [5] Kurt W. Kohn. Molecular interaction maps as information organizers and simulation guides. *Chaos*, 11(1):84–97, 2001.
- [6] Roger Brent and Larry Lok. A fishing buddy for hypothesis generators. *Science*, 308:504–506, April 2005.
- [7] Paul Thagard. Pathways to biomedical discovery. *Philosophy of Science*, 70:235–254, April 2003.
- [8] Debajit K. Biswas, Antonio P. Cruz, Eva Gansberger, and Arthur B. Pardee. Epidermal growth factor-induced nuclear factor  $\kappa$ B activation: A major pathway of cell-cycle progression in estrogen-receptor negative breast cancer cells. *PNAS*, 97(15), 2000.
- [9] Rudolf Carnap. *An Introduction to the Philosophy of Science*. Dover Publications, 1995.
- [10] David K. Lewis. *Counterfactuals*. Harvard university press, 1973.
- [11] Dongmo Zhang and Norman Foo. EPDL:a logic for causal reasoning. Technical report, University of new south wales, 2000.
- [12] Hudson Turner. Representing actions in logic programs and default theories. *Journal of Logic Programming*, 31(1-3):245–298, 1997.
- [13] A. P. Sistla and E. M. Clarke. The complexity of propositional linear temporal logics. *Journal of ACM*, 32(3):733–749, 1985.
- [14] Helen E. Bryant et al. Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase. *Nature*, 434:913–916, 2005.

- [15] Hannah Farmer et al. Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature*, 434:917–921, 2005.
- [16] Ross D. King, Kenneth E. Whelan, Ffion M. Jones, Philip G. K. Reiser, Christopher H. Bryant, Stephen H. Muggleton, Douglas B. Kell, and Stephen G. Oliver. Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, 427:247–252, January 2004.
- [17] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, , and M. Hattori. The KEGG resources for deciphering the genome. *Nucleic Acids Research*, 32:D277–D280, 2004.
- [18] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D’Eustachio, E. Schmidt, B. de Bono, B. Jassal, G.R. Gopinath, G.R. Wu, L. Matthews, S. Lewis, E. Birney, , and L. Stein. Reactome: a knowledge-base of biological pathways. *Nucleic Acids Research*, 33:D428 – D432, 2005.
- [19] Jung-jae Kim and Jong C. Park. BioIE: Retargetable information extraction and ontological annotation of biological pathways from the literature. *Journal of Bioinformatics and Computational Biology*, 2(3):551–568, 2004.