Knowledge-Based Intergrative Framework for Hypothesis Formation in Biochemical Networks

Nam Tran[1], Chitta Baral[1], Vinay J Nagaraj[2], Lokesh Joshi[2]

[1] Department of Computer Science and Engineering, Ira A. Fulton School of Engineering, Arizona State University, Tempe, AZ 85281, USA
[2] Harrington Department of Bioengineering, The Biodesign Institute at Arizona State University, Tempe, AZ 85281, USA

Abstract. The current knowledge about biochemical networks is largely incomplete. Thus biologists constantly need to revise or extend existing knowledge. These revision or extension are first formulated as theoretical hypotheses, then verified experimentally. Recently, biological data have been produced in great volumes and in diverse formats. It is a major challenge for biologists to process these data to reason about hypotheses. Many computer-aided systems have been developed to assist biologists in undertaking this challenge. The majority of the systems help in finding "pattern" in data and leave the reasoning to biologists. A few systems have tried to automate the reasoning process of hypothesis formation. These systems generate hypotheses from a knowledge base and given observations. A main drawback of these knowledge-based systems is the knowledge representation formalism they use. These formalisms are mostly monotonic and are now known to be not quite suitable for knowledge representation, especially in dealing with incomplete knowledge, which is often the case with respect to biochemical networks. We present a knowledge based framework for the general problem of hypothesis formation. The framework has been implemented by extending BioSigNet-RR. BioSigNet-RR is a knowledge based system that supports elaboration tolerant representation and non-monotonic reasoning. The main features of the extended system include: (1) seamless integration of hypothesis formation with knowledge representation and reasoning; (2) use of various resources of biological data as well as human expertise to intelligently generate hypotheses; (3) support for ranking hypotheses and for designing experiments to verify hypotheses. The extended system can be considered as a prototype of an intelligent research assistant of molecular biologists. The system is avalable at http://www.biosignet.org.

1 Introduction

Because of the complexity of living systems and the limitation of scientific methods available for the study of those systems, biological knowledge is inherently incomplete. The incompleteness of knowledge constantly manifests itself in unexplainable observations. To account for these novel observations, biologists need to revise or extend the existing knowledge. The revision and extension are first formulated as hypotheses. After being verified experimentally, a hypothesis is added to existing knowledge and becomes part of the accepted theory.

Recent advances in biological and computational sciences have produced diverse sources of biological data such as: research literature, high-throughput data (e.g.

microarray, mass spectrometry), and bioinformatic resources (e.g. interaction databases, biological ontologies). It is a major challenge for biologists to integrate these various data sets to generate hypotheses. Many computer-aided systems have been developed to assist biologists in undertaking this challenge. These systems differ in their goals, namely the automation of generating hypotheses either directly from data or based on knowledge. Although hypothesis generation from data is an important first step, often use of high-level knowledge is necessary to come of with more relevant hypothesis and to narrow down the set of hypothesis. Our work in this paper aims at contributing towards this goal.

Knowledge-based hypothesis generation has been a focus of Artificial Intelligence (AI) research in the past [36, 9]. Regarding molecular biology and in particular biochemical networks, the related works in hypothesis generation include HYPGENE [18], HinCyc [19], TRANSGENE [9], GENEPATH [42] and PathoLogic [20]. These works are built upon knowledge representation languages that are limited to "monotonic reasoning". In monotonic reasoning, if a proposition p can be concluded from a knowledge base K (denoted by $K \models p$), then p will also be concluded after K is extended with H (i.e, $K \cup H \models p$). However, the contrary is a common phenomena in biology. In that case, p becomes false after the extension of the knowledge base: $K \cup H \not\models p$. Moreover, with the exception of PathoLogic, the related works do not address the integration of multiple data sources (probably because many of the data sources were not been available at that time).

As noted above, making hypotheses from data is important because it creates the foundation to build high-level knowledge. Towards this task, a vast array of computational techniques has been developed. The computational systems produce "first-level" knowledge, which should be exploited by large-scale knowledge-based systems for hypothesis formation. It is an important requirement that such large-scale systems should allow for easy updating (referred to as "elaboration tolerance") of the knowledge base when new knowledge becomes available and avoid significant overhauling (or surgery) of the old model or scrapping of the old model and making a new model from scratch. This issue of elaboration tolerance in knowledge representation has been addressed successfully by recent advances in AI research [4].

In this work, we propose a knowledge-based framework for hypothesis formation which is based on non-monotonic reasoning and elaboration tolerant representation. We select the domain of biochemical networks as the test bed, because this domain suffers from largely incomplete knowledge and at the same time, databases and knowledge bases of biochemical networks exist in a great number. We have implemented the framework by extending the BioSigNet-RR knowledge based system [5]. We named the new system BioSigNet-RRH, which stands for "Representing, Reasoning and Hypothesizing about Biological Signal Network". Besides generating hypotheses, the new system also supports ranking of hypotheses and proposes plans for experimental verification.

The rest of the paper is organized as follows. First we discuss representative related works. Then we review basics of knowledge representation and formally define the hypothesis formation problem. We continue with the description of system and methods. Finally, we conclude with a case study of the p53 signal network.

2 Related Works

HYPGENE [18] treated the general problem of hypothesis formation as a *planning problem*. The actions are operators that modify an existing knowledge base and/or assumed initial conditions of an experiment. The goal is to resolve the mismatch between theoretical predictions computed by the knowledge base and experimental observations, with respect to the same initial conditions. The knowledge base was implemented in a frame-based representation language. HYPGENE was proposed to be domain-independent and has been tested on a problem of E.coli gene regulation. HYPGENE and BioSigNet-RRH tackle the same hypothesis formation problem that arises when an existing theory does not predict an experimental observation. The limitations of HYPGENE lie in methods, which include

- The frame-based representation language is limited to monotonic reasoning. Thus HYPGENE would have difficulty in dealing with incompleteness of biological knowledge.
- Although the biological knowledge is always incomplete, it is currently available in a great volume and in diverse formats. It is unclear how the current knowledge could have been exploited for hypothesis formation in HYPGENE.
- A hypothesis involves the modification of an existing knowledge base and/or assumed initial conditions of an experiment. HYPGENE was restricted to the modification of the initial conditions. This restricted problem amounts to a form of reasoning called *explanation* and studied in [5].

TRANSGENE [9] considered hypothesis formation as diagnosis and redesign of theories. According to this model, when a theory cannot predict an experimental observation, the theory must contain some faulty components that can be found and fixed. TRANSGENE used a "functional representation" language for knowledge representation [34]. This representation language was chosen to overcome the limitations of rule based and frame based system. Nevertheless, the language could not allow for non-monotonic reasoning. To sum up, TRANSGENE showed that limitations of knowledge representation language can seriously hinder hypothesis formation. On the other hand, it illustrates that hypothesis formation is intuitive and straightforward in knowledge based framework.

GenePath [42] automated the inference of genetic networks from experimental data. A knowledge base is a genetic network that represents positive and negative influences of a gene on another. Experiments are perturbations to the network, performed by means of gene mutations. A fixed set of inferencing rules was formalized and implemented in GenePath using Prolog. These rules encode heuristic reasoning that are routinely applied by geneticists, namely epistasis analysis. Prior background knowledge are encoded in an initial network. Starting with the initial network, GenePath applies the rules to construct a plausible network as a hypothesis that explains experimental data. GenePath can also propose new experiments for further verification and refinement of hypotheses. Although the knowledge representation and reasoning are simple in GenePath, it has illustrated the important role of expert reasoning in hypothesis formation, and that logic programming provides a straightforward and intuitive representation of human reasoning.

Integrative computational protocols [20, 26, 37] have been proposed for prediction of metabolic and regulatory pathways. They have the general scheme: (1)

construct an initial template pathway; (2) fill in missing links in the template, expand the template with new elements, or refine it; (3) verify experimentally the predicted pathway(s). These works integrated various techniques for prediction of missing genes and molecular interactions into functional contexts of pathways. They indicate that more powerful hypotheses can be found by incorporating more background knowledge and reasoning into search.

Cytoscape [35] provided an integration of various resources of molecular interaction data. By means of simulation and visualization, the system is very useful for biologist to identify novel patterns in high-throughput data. Observing novel patterns in data, biologists reason to formulate hypotheses that may explain the patterns; for example as in [3]. Cytoscape has alleviated the manual processing of high-throughput information. Nevertheless, in a near future, even the number of such patterns would also become so great that biologists would have difficult to handle such reasoning in their head. Hence, tools such as Cytoscape make the automation of reasoning to formulate hypotheses even more pressing.

HyBrow [31] was designed for computer-aided evaluation of user-defined hypotheses. A hypothesis in the HyBrow system is a set of biological events that are related logically and/or temporally. The knowledge base in HyBrow is a database integration of various data sources (e.g annotated genomic database, microarray expression data). Given a hypothesis, HyBrow checks if the hypothesis conflicts with the knowledge base. It then provides explanation for conflicts as well as suggestions for necessary refinements of the hypothesis. We will discuss later how the output of HyBrow can be useful in the hypothesis formation in BioSigNet-RRH .

Robot Scientist [21] uses machine learning techniques (active learning, decision tree, inductive logic programming) to predict gene function in metabolic networks. The knowledge representation language is a monotonic logical formalism implemented in Prolog. The system is an interesting demonstration of state-of-the-art AI methods, especially machine learning and robotics. However, it is unclear how the system can incorporate elaboration representation and non-monotonic reasoning into hypothesis formation. It is also unclear how this approach can be scaled up to take advantage of multiple sources of biological knowledge.

3 Problem Definition

Before we formally define the hypothesis formation problem, let us review some basic notions of knowledge representation.

3.1 Background of knowledge representation

In a computer system, knowledge is represented in a symbolic language with a precise syntax and semantics. For our discussion, we will use the language $\mathcal{A}_T^0$ of BioSigNet-RR [5, 39], but the general principles are applicable to any other knowledge representation formalisms.

The language $\mathcal{A}_T^0$ has an alphabet, and a restricted syntax. The alphabet of $\mathcal{A}_T^0$ consists of a set of Boolean symbols named *fluent* and a set of symbols named *action*. Fluents represent properties of the world, and actions represent mechanisms that cause the state of the world to change. For example, we can have a fluent

high(*ligand*) representing the property that the level of ligand is high. We can have an action *bind*(*ligand*, *receptor*) representing the association of ligand with receptor.

The language $\mathcal{A}_T^0$ consists of three sub-languages: a language for knowledge bases that describe the world, a language for our observations about the world, and a language for queries about the world.

A knowledge base is a set of statements in the following syntax:

$$a \textbf{ causes } f \textbf{ if } f_1, \ \ldots, \ f_k \tag{1}$$

$$g_1, \ \ldots \ , \ g_m \textbf{ triggers } b \tag{2}$$

$$h_1, \ \ldots \ , \ h_n \textbf{ inhibits } c \tag{3}$$

where a, b, c are actions, and f_i, g_j, h_k are fluents. Statements of the form (1) are called *causal rule*, which state that if a occurs in the world state s where $f_1, \ldots f_k$ are true, then f will become true in the world state s' resulted from the occurrence of a in s. Statements of the form (2) are called *trigger*, which state that action b has to occur if the preconditions $g_1, \ldots g_m$ hold. Statements of the form (3) are called *inhibition*, which state that action c cannot occur whenever the preconditions $h_1, \ldots h_n$ hold.

Example 1. Let us consider the knowledge base:

$$\begin{aligned}
&\textit{bind}(\textit{ligand}, \textit{receptor}) \textbf{ causes } \textit{bound}(\textit{ligand}, \textit{receptor}) \\
&\textit{high}(\textit{ligand}) \textbf{ triggers } \textit{bind}(\textit{ligand}, \textit{receptor}) \\
&\textit{bound}(\textit{another}, \textit{receptor}) \textbf{ inhibits } \textit{bind}(\textit{ligand}, \textit{receptor})
\end{aligned}$$

The knowledge base represents that the association of *ligand* and *receptor* results in *ligand* being bound to *receptor*; that the association occurs when the level of *ligand* is high and that the association is blocked when *receptor* is bound to another molecule. □

Observations about the world involve properties or action occurrences. To record the observation that a property f is true at time t, we write

$$f \textbf{ at } t.$$

To record the observation that some action a occurs at time t', we write

$$a \textbf{ occurs\_at } t'.$$

The semantics of $\mathcal{A}_T^0$ defines when a set $\mathcal{O}$ of observations is entailed from a knowledge base K and a set I of initial observations. The entailment is usually written as $(K, I) \models O$. For example, let K be the knowledgebase of *ligand* and *receptor*. Let I and O be the following sets of observations

$$\begin{aligned}
I =& \{\textit{high}(\textit{ligand}) \textbf{ at } 0, \neg \textit{bound}(\textit{another}, \textit{receptor}) \textbf{ at } 0\} \\
O =& \{\textit{bound}(\textit{ligand}, \textit{receptor}) \textbf{ at } 1\}
\end{aligned}$$

then $(K, I) \models O$. We also say that the observation O is explained by K, given the initial condition I.

We are now ready to discuss the general problem of hypothesis formation.

3.2 Hypothesis formation

We take the view that hypothesis formation is a reasoning process to find explanations for "novel" observations. Given a knowledge base K and initial condition I, we call an observation O "novel" with respect to K and I if O is not entailed (i.e. definitely concluded) by (K, I). For example, in the case of K and I as in the previous section, a novel observation is

$$O' = \{\neg bound(ligand, receptor) \textbf{ at } 1\}$$

With the assumption that O' is correct, we need to find explanations for O' by modifying K and I to become K' and I' such that $(K', I') \models O'$. The modification involves expansion and/or revision of the existing knowledge (i.e. K and I).

In this work, we focus on hypothesis formation as the expansion of an existing knowledge base to account for novel observations. This form of reasoning is called abduction, which was introduced by [27, 28] and has been used in various AI applications [30], including abductive logic programming [16, 17, 11, 12], diagnosis [32], planning [1, 24], default reasoning [29, 13, 16], belief revision and update [7]. We formally define hypothesis formation as follows.

Definition 1. *Let K be a knowledge base. Let O be some observation that cannot be explained by K, given some initial condition I:*

$$(K, I) \not\models O.$$

A hypothesis space is a pair $(\mathcal{S}_K, \mathcal{S}_I)$, where $\mathcal{S}_K$ is a set of rules and $\mathcal{S}_I$ is a set of observations. A hypothesis is a subset $H \subseteq \mathcal{S}_K$ such that there exists $I' \subseteq \mathcal{S}_I$ satisfying: $(K \cup H, I \cup I') \models O$. □

A hypothesis formation problem (K, I, O) is to find hypotheses as defined above.

4 System and Methods

The main steps of hypothesis formation in BioSigNet-RRH are: (1) the construction of the hypothesis space $(\mathcal{S}_K, \mathcal{S}_I)$; (2) generation of hypotheses, which includes search for and ranking of hypotheses. The ranking of hypothesis is based on estimating the complexity of each hypotheses. Simple hypotheses are preferred over complex one. Hypotheses generated by BioSigNet-RRH are theoretical and thus have to be verified experimentally. Because there are usually many ways to verify a hypothesis and biological experiments are cost sensitive, BioSigNet-RRH provides means to evaluate costs of experiments before they are performed.

We now present these major feature of BioSigNet-RRH .

4.1 Construction of hypothesis space $(\mathcal{S}_K, \mathcal{S}_I)$

In general, the rules and observations of the hypotheses space $\mathcal{S} = (\mathcal{S}_K, \mathcal{S}_I)$ include new fluent and action symbols, which form an additional alphabet. Let us denote the existing alphabet by $\mathbf{A}$ and the new alphabet by $\mathbf{A}^+$. The addition of $\mathbf{A}^+$ and the elements of $\mathcal{S}$ happen together, but we discuss them separately as follows.

Addition of $\mathbf{A}^+$. The elements of the additional alphabet $\mathbf{A}^+$come from various resources. The representative resources are as follows.

– Biologists define new fluents or actions describing biological properties or processes to be studied. There is also a wide range of techniques to infer the association between biological properties and events, for example Cytoscape [35]. If some properties and events are found to be associated with components of the knowledge base, then they should be included as fluents and actions in $\mathbf{A}^+$.
– Automated extraction of biological terms from literature has produced a great resource of biological properties and molecular interactions [33].
– Many protein interaction maps have been constructed by computational and high-throughput biological methods [40, 8]. These interaction maps can be used to define new actions.
– Biological ontologies and interaction databases [41, 2, 10] also contain biological properties and reactions as their alphabets.

Construction of $\mathcal{S}_K$. To distinguish the rules of the hypothesis space from the rules of the knowledge base, we call the former *possibilities*.

To include a possibility r in the hypothesis space, we write

$$\mathbf{POSS}[p] : r$$

where p is a non-negative number called the *preference* of r. If we do not want to take the preference into account, or if it is not available, we set $p = 0$. In the next section, we will describe how the preferences are used in ranking hypotheses.

Causal rules can be constructed from interaction databases and biological ontologies [41, 2, 10, 15]. There exists no database that contains explicit information regarding triggers and inhibitions. However, there exist datasets from which associations between properties and processes can be found. BioSigNet-RRH then takes a simple approach to generate triggers and inhibitions of the hypothesis space: if a set of fluents $f_1, f_2, \ldots f_n$ are found to be associated (or correlated) with an action a, then there are the possibilities that

$$\begin{aligned} \mathbf{POSS}[p] : \quad & f_1, f_2, \ldots f_n \ \textbf{triggers} \ a \\ & f_1, f_2, \ldots f_n \ \textbf{inhibits} \ a \end{aligned}$$

where the number p is either estimated from the data, or defined by biologists.

We can also to take advantage of data integration efforts such as HyBrow [31]. Recall that HyBrow aides in manual construction of sets of biological events that are consistent with respect to an integrated database. Such as set of events can be used as suggestions for possibilities.

Example 2. Consider a simple set of events output by HyBrow: "Gal2p transports galactose into the cell. In the cytoplasm, galactose activates Gal3p. Gal3p binds to the promoter of the Gal1 gene" [31]. Based on this set of events, there can be the following possibilities:

$$\begin{aligned} & \mathit{high}(\mathit{Gal2p}) \ \textbf{triggers} \ \mathit{trans}(\mathit{Gal2p}, \mathit{galact}) \\ & \mathit{trans}(\mathit{Gals2p}, \mathit{galact}) \ \textbf{causes} \ \mathit{in}(\mathit{galact}, \mathit{cyto}) \\ & \mathit{in}(\mathit{galact}, \mathit{cyto}) \ \textbf{triggers} \ \mathit{activates}(\mathit{Gal3p}) \\ & \mathit{activates}(\mathit{Gal3p}) \ \textbf{causes} \ \mathit{active}(\mathit{Gal3p}) \\ & \mathit{active}(\mathit{Gal3p}) \ \textbf{triggers} \ \mathit{binds}(\mathit{Gal3p}, \mathit{Gal1\_promoter}) \end{aligned}$$

Such rules are possible elements of $\mathcal{S}_K$. □

Construction of $\mathcal{S}_I$. We declare possible unknown factors in the initial conditions as follows

- f may be true or false initially: **POSS initial** f.
- a may occur initially: **POSS initial** a.

4.2 Generation of theoretical hypotheses

The reasoning in BioSigNet-RR is implemented using AnsProlog, a non-monotonic logic programming language [4]. The semantics of AnsProlog is *stable model semantics*. For example, the AnsProlog program

$$a \leftarrow not\ b$$
$$b \leftarrow not\ a$$

has 3 models $\{a\}$, $\{b\}$ and $\{a, b\}$. The models $\{a\}$ are $\{b\}$ stable, while $\{a, b\}$ is not. Stable models are minimal with respect to the $\subseteq$ ordering on sets.

The hypothesis generation in BioSigNet-RRH is also implemented using AnsProlog. A hypothesis - a set of rules - is extracted from a stable model of the AnsProlog implementation. Intuitively, we want to find hypotheses as simple as possible. The minimality of stable models has an important role towards this goal.

The ranking of hypotheses is based on the following partial ordering.

Definition 2. *Let γ be some scoring function for hypotheses. A hypothesis H is more* preferred *than a hypothesis H', written as $H \prec H'$, if $H \subset H'$ and $\gamma(H) \geq \gamma(H')$.*

A hypothesis H is *maximally* preferred, if there exists no hypothesis H' such that $H' \prec H$. We now explain how BioSigNet-RRH generates hypotheses that are maximally preferred. To ensure the minimality of hypotheses with respect to the $\subseteq$ relation search heuristics are added in the form of AnsProlog rules. Some examples of heuristics are:

- A trigger is added only if it is the only cause of some action occurrence that is needed to explain the novel observations.
- An inhibition is added only if it is the only blocker of some triggered action at some time.

The implementation of these heuristics is straightforward, and they can function as a plug-in component of BioSigNet-RRH .

The γ scoring function is currently defined such that it can be maximized using a built-in feature of the AnsProlog engine.

Let r be an element in the hypothesis space given by

$$\mathbf{POSS}[p] : r$$

Let $pref(r) = p$. The function $\gamma(H)$ is defined as the sum of the preferences of the rules in H; that is,

$$\gamma(H) = \sum_{r \in H} pref(r)$$

4.3 Guidance for experimental verification

Because of the incompleteness of biological knowledge, hypotheses can only be verified using some plausibility measure. In general, a hypothesis is accepted as a theory when there are enough experimental evidences supporting it. Thus, biologists would like to carry out as many experiments as possibile for the verification of a hypothesis. In reality, the set of possible experiments are seriously constrained by resources such as time and available techniques. Hence, it is desirable to perform only experiments that require a minimal available resource but produce a maximal information.

In this section, we propose a model of guidance for experimental verification.

Let us represent a wet-lab experiment in the abstract form (I, O), where I is the set of initial conditions of the experiment, and O is the set of observed outcomes.

Definition 3. *Let K be a knowledge base and H be a hypothesis. Let (I, O) be a experiment. We say that (I, O) is an* evidence *for the hypothesis H, if O can be explained by $K \cup H$ given I:* $(K \cup H, I) \models O$.

Example 3. Let $K = \{a \textbf{ causes } g\}$ and $H = \{f \textbf{ triggers } a\}$. Let $I_1 = \{f \textbf{ at } 0, \neg g \textbf{ at } 0\}$, $O_1 = \{g \textbf{ at } 1\}$. Let $I_2 = \{\neg f \textbf{ at } 0, \neg g \textbf{ at } 0\}$, $O_2 = \{\neg g \textbf{ at } 1\}$. Then (I_1, O_1) and (I_2, O_2) are evidences for the hypothesis H, but only (I_2, O_2) is an evidence for the hypothesis $\emptyset$. □

There are two important measures of an experiment, namely its cost and its information content. Let us denote these measure as $cost(I, O)$ and $info(I, O)$. Given a hypothesis H, the objective is to find a set E of evidences for H that has minimal cost and maximal information content. Let us simply define:

$$cost(E) = \sum_{(I,O) \in E} cost(I, O)$$

$$info(E) = \sum_{(I,O) \in E} info(I, O)$$

An initial condition such as f **at** 0 can be achieved by some wet-lab operation and can be associated with some cost. We then define

$$cost(I) = \sum_{x \in I} cost(x)$$

Biological observations are achieved by means of measurements, which also have associated costs. Hence, we define

$$cost(O) = \sum_{y \in O} cost(y)$$

Finally, $cost(I, O) = cost(I) + cost(O)$.

Let $\Omega(K, I)$ be the maximal observations that can be entailed from K, given I. That is, $(K, I) \models \Omega(K, I)$ and for all ω, if $(K, I) \models \omega$ then $\omega \subseteq \Omega(K, I)$. We define the information content of (I, O) as the deviation (or distance) of O from $\Omega(K, I)$. The distance between two sets of observations in turn is defined based on the distance between their elements.

We now present the p53 signal network as a case study to illustrate our theoretical methods to automate the process of hypothesis formation.

5 Case study

First, we describe the biology the p53 network in parallel with its knowledge-based representation.

5.1 p53 signal network

The p53 protein plays a central role as a tumor suppressor and is subjected to tight control through a complex mechanism involving several proteins. The key aspects of the p53 network are as follows.

Tumor suppression by p53: The p53 protein has three main functional domains; the N terminal transactivator domain, the central DNA-binding domain and a C terminal domain that recognizes DNA damage. The binding of the transactivator domain to the the promoters of target genes activates pathways to lead to a reversible arrest of the cell cycle, prevention of genomic instability or apoptosis and thus protects the cell from cancer [23]. The ability to suppress tumors is retained when the interacting partners of p53 do not inhibit the functionality of the transactivator domain.

fluent $bound(dom(p53, N))$
action $grow(tumor)$
$high(p53)$ **inhibits** $grow(tumor)$
$high([p53 : P]), not\ bound(dom(p53, N))$ **inhibits** $grow(tumor)$

(The keywords **fluent** and **action** are used to declare fluent and action symbols in BioSigNet).

Interaction between Mdm2 and p53: Mdm2 binds to the transactivator domain of p53, thus inhibiting the p53 induced tumor suppression. The binding of Mdm2 to p53 also causes changes in the protein concentration levels.

fluent $high(p53), high(mdm2), high([p53 : mdm2])$
action $bind(p53, mdm2)$
$bind([p53 : mdm2])$ **causes** $bound(dom(p53, N))$
$high(p53), high(mdm2)$ **triggers** $bind(p53, mdm2)$
$bind(p53, mdm2)$ **causes** $high([p53 : mdm2])$,
$bind(p53, mdm2)$ **causes** $\neg high(p53), \neg high(mdm2)$

Mdm2 induced degradation of p53: Under normal physiological conditions, p53 levels remain low due to rapid and constant turnover. The short half life of p53 is due to the formation of a complex with Mdm2 that gets targeted for ubiquitin dependent proteosomal degradation.

action $degrade(p53, mdm2)$
$high([p53 : mdm2])$ **triggers** $degrade(p53, mdm2)$
$degrade(p53, mdm2)$ **causes** $\neg high([p53 : mdm2])$

Upregulation of p53: The elevated levels of p53 may be a result of upregulation of p53 gene expression, increased transcript stability, enhanced translation of p53

mRNA [14], or post-translational modifications of the p53 protein which favor a prolonged half life and increased activity [6].
For the case study, we consider the upregulation of p53 expression, which is represented as follows.

$$upregulate(mRNA(p53)) \textbf{ causes } high(mRNA(p53))$$
$$high(mRNA(p53)) \textbf{ triggers } translate(p53)$$
$$translate(p53) \textbf{ causes } high(p53)$$

Stress: UV, ionizing radiation, and chemical carcinogens cause stress. Stress can induce the upregulation of p53.

$$high(UV) \textbf{ triggers } upregulate(mRNA(p53))$$

Stress can induce changes in expression of tumor related genes, (e.g. cmyc), which result in uncontrolled cell division (tumor).

$$high(UV) \textbf{ triggers } alter(expr(cmyc))$$
$$alter(expr(tumorgenes)) \textbf{ causes } altered(expr(cmyc))$$
$$altered(expr(cmyc)) \textbf{ triggers } grow(tumor)$$
$$grow(tumor) \textbf{ causes } tumorous$$

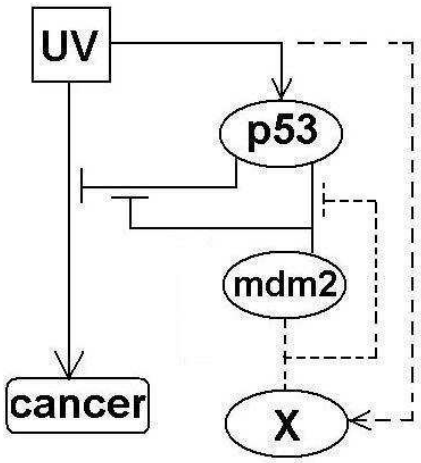

Fig. 1. A hypothesis in p53 interaction network. The $\rightarrow$ represents trigger. The $\dashv$ represents inhibition. The solid and dash lines represent known and hypothetical interactions, respectively.

Given the theory of the p53 network, a hypothesis formation problem arises as follows.

5.2 The problem

X is a tumorsuppressor gene. Mutants of X are highly susceptible to cancer. We would like to hypothesize on the various possible influences of X on the p53 pathway.

Thus, we have the hypothesis problem (K, I, O), where K is the knowledge base of p53 biology, and I is the initial condition

$$I = \{null(X) \textbf{ at } 0\}$$

and O is the observation

$$O = \{\textbf{eventually } tumorous\}$$

(Here, **eventually** F is a logical proposition denoting that some property F will be true at some future time).
We need to extend K with H such that there exists I' satisfying: $(K \cup H, I \cup I') \models O$.

5.3 Hypothesis formation

Construction of the hypothesis space First, we show how various possibilities can be found and included in the hypothesis space. In the following, the literature means [23, 14, 6].

There may be functional similarities between X and p53: X is a tumor suppressor, so we have a prior knowledge that X may play the same effects as p53 in stressed cells, which is described in the following possibilities:

$$\begin{aligned}\textbf{POSS}: \quad & high(UV) \textbf{ triggers } upregulate(mRNA(X)) \\ & upregulate(mRN(X)) \textbf{ causes } high(mRNA(X)) \\ & high(mRNA(X)) \textbf{ triggers } translate(X) \\ & translate(X) \textbf{ causes } high(X)\end{aligned}$$

Stress may induce high level of X: Data from the literature show that the levels of protein X is found to be higher in cells subjected to stress. Consequently, it is possible that stress induces the upregulation of X expression. That is,

$$\textbf{POSS}: \quad high(UV) \textbf{ triggers } upregulate(mRNA(X))$$

X or p53 may induce upregulation of the other: There are observations from the literature that high levels of X are concomitant with elevated levels of p53. Thus, it is possible that a high level of X induces the upregulation of p53, or vice versus.

$$\begin{aligned}\textbf{POSS}: \quad & high(X) \textbf{ triggers } upregulate(mRNA(p53)) \\ & high(p53) \textbf{ triggers } upregulate(mRNA(X))\end{aligned}$$

X may interact with the known proteins in the network: The possible interactions are $bind(p53, X)$ and $bind(mdm2, X)$. The possibile properties are the protein levels and the domains of p53. By associating a possible action with possible effects, the system automatically includes the possibilities such as

$$\begin{aligned}\textbf{POSS}: \quad & bind(p53, X) \textbf{ causes } bound(dom(p53, N)) \\ & bind(p53, X) \textbf{ causes } \neg bound(dom(p53, N))\end{aligned}$$

That is, binding of X to p53 may or may not affecting the transactivator domain.

X may influence (trigger/inhibit) other interactions: The system automatically includes all the possibilities of X's influences on the interactions in the network, resulting in

$$\begin{aligned} \textbf{POSS}: \quad & high(X) \ \textbf{influences} \ upreg(mRNA(p53)) \\ & high(X) \ \textbf{influences} \ translate(p53) \\ & high(X) \ \textbf{influences} \ bind(p53, mdm2) \end{aligned}$$

(where **influences** stands for either **triggers** or **inhibits**).

Hypotheses generation We present representative examples of the hypotheses generated by BioSigNet-RRH .

- *X is a negative regulator of Mdm2:* Stress induces high expression of X. X binds to Mdm2 and this complex is rapidly degraded by proteolysis. Scavenging of Mdm2 arrests the proteolyis p53 (Fig. 1). The important elements of the hypothesis are:

$$\begin{aligned} & high(UV) \ \textbf{triggers} \ upregulate(mRNA(X)) \\ & high(X), high(mdm2) \ \textbf{triggers} \ bind(X, mdm2) \end{aligned}$$

- *X directly influences p53 protein stability:* X binds to p53 protein at a domain different from the transactivator domain, so p53 is stabilized (formation of Mdm2-p53 complex is prevented) and still functional as tumor suppressor. The important elements of the hypothesis are:

$$\begin{aligned} & high(X), high(p53) \ \textbf{triggers} \ bind(p53, X) \\ & bind(p53, X) \ \textbf{causes} \ \neg bound(dom(p53, N)) \end{aligned}$$

The non-monotonicity of the framework manifests itself in the results. The knowledge base in Section 5.1 predicts that cancer will finally occur due to high level of UV (stress). After being extended with the hypothesis described in Fig. 1., the new knowledge base predicts that cancer will not occur, given the presence of UV.

The presented study is incomplete in the sense that changes in the regulation of p53 also occurs as a result of stress induced damage to DNA. Due to the elaboration tolerance feature, we could start by first constructing a small initial knowledge base, then incrementally adding more knowledge. We have also represented simple rules with only one or two preconditions. More elaborated representation and the results on experiments with ranking can be found at the system's Website.

6 Conclusion

We have presented a general framework for the automation of hypothesis formation in systems biology. We considered the hypothesis formation problem in the context of knowledge representation and reasoning. We implemented an initial system by extending BioSigNet-RR. The advantages of our approach includes: (1) hypothesis formation is defined as a form of reasoning and is implemented using AnsProlog, which is an elaboration tolerant and non-monotonic representation and reasoning

language; (2) it provides a mean to integrate various resources of biological knowledge; (3) it is a high-level approach to hypothesis formation that is necessary for building an intelligent system to aid biologists.

Our work is a proof-of-concept and substantial works remain for the scaling-up the system for real-world applications. We identify many important future works. First, it is important to allow for declaration and instantiation of "similarity" background knowledge; such as gene homology, or the similarity between relationships between proteins or biological processes. Next, we want to explore different models of model ranking. We will explore how AnsProlog with preferences can be applied for model ranking. Towards this goal, we plan to take advantage of the large body of research results on answer set programming with preferences. Finally, we have restricted to the hypothesis formation as knowledge extension. Hypothesis formation based on knowledge revision is an important next development.

Bibliography

[1] J. Allen, H. Kautz, R. Pelavin, and J. Tenenberg. *Reasoning about plans*. Morgan Kaufmann, San Mateo,CA, 1991.

[2] G. D. Bader, I. Donaldson, C. Wolting, B. F. F. Ouellette, T. Pawson, and C. W. V. Hogue. Bind the biomolecular interaction network database. *Nucleic Acids Research*, 29(1):242–245, 2001.

[3] N. S. Baliga, S. J. Bjork, R. Bonneau, M. Pan, C. Iloanusi, M. C. Kottemann, L. Hood, and J. DiRuggiero. Systems Level Insights Into the Stress Response to UV Radiation in the Halophilic Archaeon Halobacterium NRC-1. *Genome Res.*, 14(6):1025–1035, 2004.

[4] C. Baral. *Knowledge representation, reasoning and declarative problem solving*. Cambridge University Press, 2003.

[5] C. Baral, K. Chancellor, N. Tran, N. Tran, and M. Berens. A knowledge based approach for representing and reasoning about signaling networks. *Bioinformatics 20 (Suppl 1)*, pages i15–i22, 2004.

[6] A. M. Bode and Z. Dong. Post-translational modification of p53 in tumorigenesis. *Nat. Rev. Cancer.*, 4(10):793–805, 2004.

[7] C. Boutilier. Abduction to plausible causes: An even based model of belief update. *Artificial Intelligence*, 83:143–166, 1996.

[8] T. Bouwmeester and et. al. A physical and functional map of the human TNF-alpha/NF-kappaB signal transduction pathway. *Nat. Cell. Biol.*, 6(2):97–105, 2004.

[9] L. Darden. Recent work in computational scientific discovery. In *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society*, pages 161–166, 1997.

[10] E. Demir, O. Babur, U. Dogrusoz, A. Gursoy, A. Ayaz, G. Gulesir, G. Nisanci, and R. Cetin-Atalay. An ontology for collaborative construction and analysis of cellular pathways. *Bioinformatics*, 20(3):349–356, 2004.

[11] M. Denecker and A. C. Kakas. Abduction in Logic Programming. In *Computational Logic: Logic Programming and Beyond*, pages 402–436, 2002.

[12] P. Doherty, S. Kertes, M. Magnusson, and A. Szalas. Towards a Logical Analysis of Biochemical Pathways. In *Proc. of JELIA*, 2004.

[13] K. Eshghi and R. Kowalski. Abduction compated with negation as failure. In *Proc. 6th Inter. Conf. in Logic Programming*, pages 234–255, 1989.

[14] T. Hamid and S. Kakar. PTTG/securin activates expression of p53 and modulates its function. *Mol. Cancer.*, 3(1):18, 2004.

[15] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. Gopinath, G. Wu, L. Matthews, S. Lewis, E. Birney, and L. Stein. Reactome: a knowledgebase of biological pathways. *Nucl. Acids Res.*, 33(Suppl_1):D428–432, 2005.

[16] A. Kakas, R. Kowalski, and F. Toni. The role of abduction in logic programming. *Handbook of logic in Artificial Intelligence and Logic Programming*, pages 235–324, 1998.

[17] C. Kakas, Antonis, B. Van Nuffelen, and M. Denecker. A-system : Problem solving through abduction. In *Proc. of the IJCAI*, volume 1, pages 591–596, 2001.

[18] P. D. Karp. Design methods for scientific hypothesis formation and their application to molecular biology. *Machine Learning*, 12:89–116, 1993.

[19] P. D. Karp, C. Ouzounis, and S. Paley. HinCyc: A Knowledge Base of the Complete Genome and Metabolic Pathways of H. influenzae. In *Proc. of ISMB*, 1996.

[20] P. D. Karp, S. Paley, and P. Romero. The pathway tools software. *Bioinformatics*, 18(Suppl. 1):S225–S232, 2002.

[21] R. King and et. al. Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, 427(6971):247–52, 2004.

[22] A. Krause, J. Stoye, and M. Vingron. The SYSTERS protein sequence cluster set. *Nucleic Acids Research*, 28(1):270–272, 2000.

[23] D. Michael and M. Oren. The p53 and Mdm2 families in cancer. *Curr. Opin. Genet. Dev.*, 12(1):53–59, 2002.

[24] M. Missiaen, L. Bruynooghe, and M. Denecker. CHICA: A planning system based on event calculus. *J. Logic Comput.*, 5(5):579–602, 1995.

[25] N. Mulder and et.al. The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Research*, 31:315–318, 2003.

[26] A. Osterman and R. Overbeek. Missing genes in metabolic pathways: a comparative genomics approach. *Current Opinion in Chemical Biology*, 7:238–251, 2003.

[27] C. Peirce. *Collected papers of Charles Sanders Peirce, Vol. 1-8*. Havard University Press, Cambridge, MA, 1931-1958.

[28] C. Peirce. *Reasoning and the Logic of Things*. Havard University Press, Cambridge, MA, 1992.

[29] D. Poole. A logical framework for default reasoning. *Artificial Intelligence*, 36(1):27–48, 1988.

[30] D. Poole, A. Mackworth, and R. Goebel. *Computational Intelligence*. Oxford University Press, Oxford, 1998.

[31] S. A. Racunas, N. H. Shah, I. Albert, and N. V. Fedoroff. HyBrow: a prototype system for computer-aided hypothesis evaluation. *Bioinformatics*, 20(Suppl_1):i257–264, 2004.

[32] R. Reiter. A theory of diagnosis from first principles. *Artificial Intelligence*, 13(1–2):81–132, 1980.

[33] A. Rzhetsky, I. Iossifov, T. Koike, M. Krauthammer, P. Kra, M. Morris, H. Yu, P. A. Duboue, W. Weng, W. J. Wilbur, V. Hatzivassiloglou, and C. Friedman. Geneways: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *J. of Biomedical Informatics*, 37(1):43–53, 2004.

[34] V. Sembugamoorthy and B. Chandrasekaran. Functional Representation of Devices and Compilation of Diagnostic Problem-Solving Systems. *Experience, Memory and Reasoning*, pages 47–73, 1986.

[35] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, 13(11):2498–2504, 2003.

[36] J. Shrager and P. Langley. *Computational Models of Scientific Discovery and Theory Formation*. Morgan Kaufmann, 1990.

[37] Z. Su, P. Dam, X. Chen, V. Olman, T. Jiang, B. Palenik, and Y. Xu. Computational inference of regulatory pathways in microbes: an application to phosphorus assimilation pathways in synechococcus sp. wh8102. In M. Gribskov, M. Kanehisa, S. Miyano, and T. Takagi, editors, *Genome Informatics*, volume 14, pages 3–13, 2003.

[38] R. Tatusov and et.al. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4(1), 2003.

[39] N. Tran and C. Baral. Reasoning about triggered actions in AnsProlog and its application to molecular interactions in cells. In *Proc. of KR 2004*, pages 554–563, 2004.

[40] A. Valencia and F. Pazos. Computational methods for the prediction of protein interactions. *Curr Opin Struct Biol.*, 12(3):368–73, 2003.

[41] I. Xenarios, D. E. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte, and D. Eisenberg. Dip: The database of interacting proteins. *Nucleic Acids Research*, 28(1):289–91, 2000.

[42] B. Zupan and et al. Genepath: a system for inference of genetic networks and proposal of genetic experiments. *Artif. Intell. Med.*, 29(1-2):107–30, 2003.