1

Symbolic Representation and Inference of Regulatory Network Structures

Recent results have demonstrated the usefulness of symbolic approaches for addressing various problems in systems biology. One of the fundamental challenges in systems biology is the extraction of integrated signaling-transcriptional networks from experimental data. In this chapter, we present a general logic-based framework, called Abductive Regulatory Network Inference (ARNI), where we formalize the network extraction problem as an abductive inference problem. A general logical model is provided that integrates prior knowledge on molecular interactions and other information for capturing signal-propagation principles and compatibility with experimental data. Solutions to our abductive inference problem define signed-directed networks that explain how genes are affected during the experiments. Using in-silico datasets provided by the dialogue for reverse engineering assessments and methods (DREAM)) consortium, we demonstrate the improved predictive power and complexity of our inferred network topologies compared with those generated by other non-symbolic inference approaches, showing the suitability of our approach for computing complete realistic networks. We also explore how the improved expressiveness together with the modularity and flexibility of the logic-based nature of our approach can support automated scientific discovery where the validity of hypothesized biological ideas can be examined and tested outside the laboratory.

1.1. Introduction: logical modeling and abductive inference in systems biology

Systems biology is generally concerned with developing formal models that aim to describe the operation of various biological processes. Its study is based on the

Chapter written by Nataly MAIMARI, Krysia BRODA, Antonis KAKAS, Rob KRAMS and Alessandra RUSSO.

synthesis of a model or a theory from empirical experimental information. At the cellular level, systems biology aims to build models that describe, at some level of abstraction, the underlying operation of a cell at the genomic and/or protein level. The central challenge is then how to choose an appropriate framework that would (1) enable the construction of a model from experimental data and (2) empower such models with a predictive capability for new information beyond the one used to construct the model.

As in many cases of such scientific exploration, the choice of the framework under which we formulate the model depends on the type of experimental data that is available at the time of the development of the scientific model. In general, at the initial stages of an investigation the available data is usually descriptive and qualitative rather than quantitative. As such we set out to develop a first model, based on some principles that we believe underlie the phenomena, where we are primarily interested in capturing the overall and general interrelation between the concepts of interest. It is then important to require a framework that is (1) high-level close to the human description of the phenomena and thus close to the experimental language, and (2) modular and flexible so that the models can easily be adapted to new information and other changes that might come about.

Under these conditions and requirements for our language, a *symbolic or logical* framework is particularly suitable. A logical scientific theory normally offers a high-level declarative description that can be understood easily by the expert experimental scientists that provide the experimental data. Logical models are also highly modular where changes can often be isolated to parts of the model without the need for an overall complete reformulation of the model. Furthermore, within a logical approach we can employ *abductive reasoning* to help in the process of building a theory from experimental data. Abductive reasoning is a formalization of the explanatory scientific reasoning that is typically carried out by human scientists when they think about the phenomena they are studying, either when they are trying to understand their experimental findings, or when they are planning the next set of experiments to help them improve their understanding of the phenomena.

Hence, in choosing a logical approach, we provide a framework that not only responds well to the object level requirement of describing the phenomena, but also to the meta level task of reasoning about the models developed thus far and deciding on their further investigation through new experiments, or indeed new desirable properties and principles that the model must adhere to. For molecular biology, logic is particularly suited as, at least currently, in many cases the theoretical models and experimentation of cell biology are developed following a rationale at the qualitative rather than quantitative level. The nature of much of the experimental data is descriptive with the aim to first understand the qualitative interrelations between the various constituents and processes in the cell.

In this chapter, we have developed a logical model of regulatory cell networks, covering both transcriptional networks and upstream signaling regulatory networks. We have implemented a qualitative model that is based on general biological principles and which exploits current prior knowledge of molecular interactions that are already known. The approach, called *ARNI*, for abductive inference of regulatory networks, constructs causal signed-directed networks of interactions between genes from high-throughput experimental data. These networks rely on the simple and general underlying principles that signals from the environment propagate along paths of protein interactions to reach the regulatory components of cells (i.e. production of genes) and that genes are under the influence of multiple overlapping inputs, which might be compatible or competitive to each other. The networks also exhibit several important motifs including feedback loops (positive and negative), which allow a gene to control its own expression, and feed-forward loops (coherent or incoherent), whereby a gene has both direct and indirect connections to its target[1]. Each of these motifs governs fundamental properties of the overall dynamic behaviorof the network such as robustness, oscillations, memory and bistability [ALO 07, YEG 04].

Our construction of regulatory networks relies on abductive reasoning as an automated form of the scientific reasoning of rationalizing the high throughput experimental data. Indeed, the problem of signaling network reconstruction naturally maps to an abductive task. Specifically, (1) gene expression data constitutes the experimental data; (2) the given (partial) knowledge is a logic-based theory governing biological phenomena, as for instance the notions of gene regulation, interactive potential; (3) biological constraints like sign consistency between interacting gene expressions are captured via integrity constraints and (4) sentences about unknown compatible and competitive gene regulations are the abducible information that can be assumed to form a network. Thus, assuming the general possible structure of signaling networks an abductive computation results in the inference of possible signed-directed networks, in terms of compatible and competitive gene regulations, that conform to the available experimental observations.

As argued above, our logical approach offers a high-level declarative model with suitable and increased expressiveness for the wide applicability to a variety of signaling network problems and challenges. We demonstrate these properties of the approach through a series of evaluation experiments that test the effectiveness of the abductive networks and explore the expressiveness of the logical framework. We also examine the usefulness of our abductive approach in the meta-level scientific reasoning, as a scientific assistant and how this, together with the modularity of the

1 Feed forward loop motifs are either coherent, if the direct effect of the regulator is the same at its net indirect effect, or incoherent otherwise.

approach, can support the further development and improvement of the initially constructed networks.

Our approach follows a series of works that rely on logical abduction for addressing various problems in systems biology. Abduction has been used to learn/revise metabolic pathways [RAY 10, TAM 06] and to hypothesize on the function of genes [RAY 08, KIN 04]. Abductive reasoning is also used in [TRA 09, LAZ 13] for meta-level reasoning over hypotheses but de-novo topology inference is not considered in these existing contributions. More directly related to our work is the approach in [PAP 05], which uses abductive logic programming to infer gene dependencies to explain the changes in the gene expression levels. Our work advances that in [PAP 05] in several ways, specifically by allowing the use of prior knowledge, modeling and reasoning about competitive gene influences and presenting a framework that can act as a scientific assistant to biologists for testing the validity of new hypotheses.

In comparison with non-symbolic approaches such as gene co-expression networks based on statistical principles [ROT 13, HE 09] and physical network models [YEA 04, OUR 07, HUA 09], logical approaches like ours offer improved expressiveness, as they enable the inference of networks with more complex regulatory structures, and added modularity that allows the logic model to be easily adapted to new available information (e.g. addition of new constraints).

This chapter is structured as follows. Section 1.2 presents the ARNI approach with its main key components. Section 1.3 describes the results on evaluating the predictive power of our approach and demonstrates the increased expressive power of ARNI. Section 1.4 explores ARNI as a scientific assistant for biological hypothesis testing and section 1.5 concludes the chapter with a discussion on related work and future directions.

1.2. Logical modeling of regulatory networks

In this section, after briefly summarizing the basic notions and terminology from abduction, we study how the problem of inferring regulatory networks can be formalized as an abductive problem. We analyze the general biological features of the problem and develop the underlying logical model over which the task of constructing regulatory networks from experimental data can be understood and computationally realized in terms of abduction.

1.2.1. *Background*

An *atomic* formula (or *atom* in brief) is a proposition or an n-ary predicate P followed by an n-tuple of terms.A *positive* literal is an atom ϕ, and a *negative* literal

is a negated atom, written as $\texttt{not}\ \phi$, where $\texttt{not}$ is the negation as failure operator. Positive or negative atoms are referred to as *literals*. A *clause* is a *rule* of the form $\phi \leftarrow \phi_1, \ldots, \phi_n$, where ϕ is the *head* atom and ϕ_i are the body literals. Clauses can also be *facts* (when $n = 0$), or *denials* of the form $\texttt{ic} \leftarrow \phi_1, \phi_2, \ldots, \phi_{\texttt{m}}$, where the symbol $\texttt{ic}$ means $false$ and ϕ_i are literals. A clause is said to be *ground* if it contains no variables, *definite* if all its body literals are positive, and *normal* if it includes at least one negative body literal. A *normal* logic program is a set of normal clauses. In general, a *model* I of a set Π of normal clauses, is a set of ground atoms such that, for each ground instance r_g of a clause r in Π, I satisfies the head of r_g whenever it satisfies the body. A model I is said to be $minimal$ if it does not strictly include (in terms of set inclusion) any other model. Normal logic programs may have one, none, or several minimal models. It is usual to identify these minimal models, called *stable models*, as the possible meanings of a program [GEL 88].

Abduction is a process of reasoning from observations to possible causes. In essence, it is concerned with the construction of explanations, Δ, that conform with given observations and prior knowledge, Π, and that, together with Π, are consistent with given integrity constraints, IC. Abductive explanations are usually restricted to ground atoms from a predefined set called *abducibles*. Intuitively, abducibles are undefined information in a given knowledge base, whose truth value can be assumed to (partially) complete the knowledge base. In logic terms, given a set Π of normal clauses, expressing prior knowledge and observations, a set IC of denials, and a set $\mathcal{A}$ of abducible ground atoms, with terms from the Herbrand domain of Π, an abductive reasoning problem consists of finding a set of abducibles $\Delta \subseteq \mathcal{A}$ such that IC is satisfied in a canonical model of $\Pi \cup \Delta$. We assume as canonical models, the *stable models* of $\Pi \cup \Delta$. Such stable models are also referred to as *generalized stable models* of the abductive task [KAK 90].

DEFINITION 1.1.– *Let the tuple $AC = \langle \Pi, IC, \mathcal{A} \rangle$, be an abductive problem, where Π is a normal logic program, IC is a set of denial clauses, and $\mathcal{A}$ is a set of ground abducible atoms. A generalized stable model of AC is a stable model of $\Pi \cup \Delta$ for some $\Delta \subseteq \mathcal{A}$ that satisfies the IC, denoted $\Pi \cup \Delta \models IC$. The set Δ is referred to as an* abductive solution *of AC.*

Different abductive proof procedures have been proposed (e.g. [KAK 90, KAK 00, KAK 01]). In these approaches, a *minimality* criterion, expressed in terms of subset-minimality, is often enforced on the construction of abductive solutions. But, whereas minimality of explanations is desirable in applications of abduction such as planning and diagnosis, extracting regulatory networks that conform with observed gene expression data means computing *maximal* networks that are biological meaningful (i.e. satisfy biological integrity constraints), that are consistent with prior knowledge about the observed genes (e.g. existing knowledge of a gene being an activator or an inhibitor), and that, together with the prior knowledge, satisfies the observed data. The computation of any such

network, in terms of collection of regulations between genes (i.e. *compatible* or *competitive* gene regulations), would require an abductive task for which abductive solutions (i.e. the regulations between genes) are not minimal but in fact *maximal*.

The answer set programming (ASP) paradigm provides the ideal environment for efficient computation of maximal abductive solutions, as it combines a declarative modeling language with high-performance problem solving computational capabilities [GEB 12]. To understand how an abductive problem, with prior knowledge Π, abducibles $\mathcal{A}$ and integrity constraints IC, is modeled in terms of an ASP problem, it is easy to think of it as a special type of open program, $\langle \Pi \cup IC, \emptyset, \mathcal{A} \rangle$ [BON 01] where the set $\mathcal{A}$ of open predicates (i.e. predicates that are not defined in the program) is the set of abducibles, and $\emptyset$ denotes that no new terms, in addition to those in the Herbrand base of Π, are considered in $\mathcal{A}$ [BON 02]. Abducibles can indeed be seen as ground Boolean atoms whose truth value is not defined in the program Π, although it is constrained by IC. In biological terms, our (abductive) problem of extracting genes regulatory network assumes that information about regulations between genes (i.e. compatible regulation or competitive regulations), which are our abducibles, is unknown and therefore "open" to Boolean assignments. Open programs can be transformed into semantically equivalent normal logic program representations (see [BON 02] for a precise definition of such semantical equivalence), which, in turn, can be expressed as ASP problems with a *choice statement* over subsets of $\mathcal{A}$ (see [GEB 12] for the mapping between normal logic programs and choice statements). A choice statement is an expression of the form $\{\mathtt{a}_1, \mathtt{a}_2, \ldots, \mathtt{a}_m\}$, where a_i are (possibly ground) atoms. This expression informally means that a subset of $\{\mathtt{a}_1, \mathtt{a}_2, \ldots, \mathtt{a}_m\}$ is included in a stable model (i.e. answer set solution) of the given ASP problem. As the set of ground abducibles could be large, choice statements can be expressed more concisely using *conditional literals* [GEB 12]. Conditional literals are expressions of the form $\mathtt{a} : \mathtt{t}_1 : \ldots : \mathtt{t}_n$, where $\mathtt{a}$ and $\mathtt{t}_i$ are literals, informally denoting the list of elements in the set $\{a \mid t_1, \ldots, t_n\}$. Clearly the expansion of conditional literals is domain dependent, i.e. it depends on the definition of the literals $\mathtt{t}_i$. So, for example, given the following literals $\mathtt{p}(1)$, $\mathtt{p}(2)$, $\mathtt{p}(3)$ and $\mathtt{q}(2)$, a choice statement $\{\mathtt{r(1)}, \mathtt{r(3)}\}$ could also be written as $\{\mathtt{r(X)} : \mathtt{p(X)}, \mathtt{not\ q(X)}\}$.

The formalization of an abductive problem in terms of an ASP problem allows better control on the size of the subset of abducibles that can be included in a final solution, taking also into account different weights that could be given to different abducibles (if required by the problem domain). For instance, we may want to specify that a solution (i.e. answer set) should include the maximal (respectively, minimal) number of abducibles that are consistent with the prior knowledge and the integrity constraints. An ASP problem would in this case include, together with the prior knowledge and the integrity constraints, the *optimization* expression `maximize` (respectively, `minimize`) over the set of the abducibles. An optimization expression is of the form $\mathtt{minimize}\{\mathtt{l}_1 = \mathtt{w}_1@\mathtt{p}_1, \ldots, \mathtt{l}_n = \mathtt{w}_n@\mathtt{p}_n\}$, and similarly for the

case of maximize but with the term `minimize` replaced by `maximize`, where $\mathtt{w}_i$ and $\mathtt{p}_i$ represent the weight and the priority of the literal $\mathtt{l}_i$. Informally, optimization expressions are directives to instruct an ASP solver to compute optimal stable models by minimizing (or maximizing) a weighted sum of elements. It is easy to see that using a maximize expression for the choice of subsets over the set of abducibles, and assuming that each abducible has the same weight and same priority (i.e. $\mathtt{maximize}\{\mathtt{a}_1, \mathtt{a}_2, \ldots, \mathtt{a}_m\}$), we basically model the requirement that optimal solutions (i.e. stable models) will include maximal number of abducibles (in terms of set inclusion). The satisfiability of the integrity constraints will be implicitly guaranteed by the computation of the optimal stable models as the ASP problem directly includes the IC.

To analyze further the difference between our emphasis on maximality versus the more conventional notion of minimality of abductive solutions, and its biological relevance in the computation of regulatory networks, we consider a simple illustrative example. Suppose that our abductive task is to compute an acyclic directed graph with four nodes a, b, c and d that links two of these nodes, say a and b, called *seed* nodes, by passing through the other two nodes c and d and satisfying the following constraints: (1) seed nodes cannot be linked directly, (2) any two nodes can have at most one link between them, (3) a seed node can either be a *source* (i.e. its links are all directed out), or a *sink* (i.e. its links are all directed in) and (4) no other node is a source or a sink (i.e. if a link exists from node Y to node X, then there must exist a link directed out from node X and a link directed into node Y). Essentially, constraint (4) guarantees the formation of paths between seed genes. We show how this abductive problem is formalized within the ASP paradigm and discuss differences between minimal and maximal solutions.

Figure 1.1 shows the ASP formalization of our abductive task $\langle \Pi, IC, \mathcal{A} \rangle$. It can be shown that this representation corresponds to a normal logic program transformation of an open program $\langle \Pi \cup IC, \emptyset, \mathcal{A} \rangle$. The ASP problem in Figure 1.1 returns many answer set solutions corresponding to different possible subsets (including the empty set) $\Delta \subset \mathcal{A}$ that are consistent with the constraints. These are determined by means of the choice statement $\{\mathtt{r(X,Y) : node(X) : node(Y) : X != Y}\}$. So, in this example, abductive solutions are finite sets of ground instances of `r(X,Y)`, i.e. directed links between nodes, that satisfy the constraints (i)-(iv). To compute just solutions that have minimal abductive assumptions, the above ASP problem can be augmented with the optimization expression $\sharp\mathtt{minimize}\{\mathtt{r(X,Y)}\}$. Clearly, in this case, the smallest set of abducibles that satisfy constraints (1)-(4) is the empty set, and the solver will return the solution with $\Delta = \emptyset$ as optimal solution. We could consider the addition of constraints to force as many links as possible to be abduced. For instance, constraint (5) *every node must be linked in the graph* could be added to the set of ICs by including the two denials `:-node(X),not connected_out(X).` and `:-node(X), not connected_in(X).` The empty solution would in this case

not be computed, as it would violate constraint (v); but the minimize optimization statement would generate, as optimal, all possible solutions satisfying all constraints that guarantee all nodes to be connected but with the minimum number of links. The abductive problem accepts in this case four minimal abductive solutions, which are graphically given in Figure 1.2, where an arrow between two nodes (e.g., d and c) represents a ground abduced r atom (e.g., $r(c,d)$). Although logically correct, such solutions are not biologically very meaningful. In real biological networks, genes (nodes in the graph) are often involved in multiple interactions (i.e. multiple incoming links or multiple outgoing links). This redundant structure of parallel overlapping inputs, ensures robustness under random failure and provides adaptability to the environment [BAR 04].

Background Π	Integrity Constraints IC	Abducibles $\mathcal{A}$
node(a). node(b). node(c). node(d). seed(a). seed(b). %connected_out(X) when link out of X %connected_in(X) when link into X connected_out(Z):- r(X,Z). connected_in(X):- r(X,Z). % special case for seed nodes connected_in(X):- seed(X). connected_out(X):- seed(X).	% constraint (i) :-r(X,Y), seed(X), seed(Y). % constraint (ii) :-r(X,Y), r(Y,X). % constraint (iii) :-r(X,Y), r(Z,X), node(Y), node(Z), seed(X). % constraint (iv) :-r(X,Y), not connected_out(X). :-r(X,Y), not connected_in(Y).	{r(X,Y):node(X):node(Y): X!=Y }.

Figure 1.1. *An abductive task as an ASP problem*

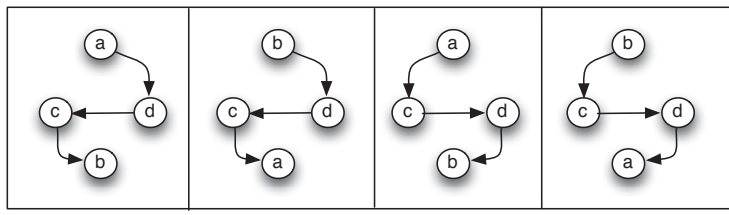

Figure 1.2. *Minimal abductive solutions that satisfy constraints (1)-(5)*

What we need in our problem is to compute maximal networks. This is achieved by requiring abductive solutions to be maximal. By adding to the same ASP problem in Figure 1.1 the constraint (v) described above and the optimization expression `#maximize{r(X,Y)}` over the choice of subset of abducibles, the abductive problem would have, in this case, still four solutions but maximal. The solutions are graphically described in Figure 1.3.

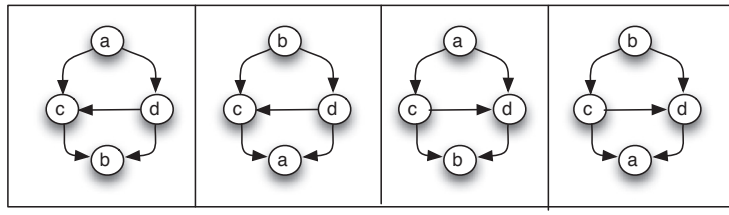

Figure 1.3. *Maximal abductive solutions for abductive task in Figure 1.1*

In summary, the task of computing regulatory networks from gene expression data can be formalized as an abductive task where maximal abductive solutions are computed to give maximal signed-directed gene regulations that are consistent with biological constraints and given gene expression data.

1.2.2. *Logical model of signed-directed networks*

In our ARNI abductive framework, the background knowledge Π is composed of a rule-based model, called *formal model*, an extensional knowledge, called *prior knowledge*, and information about experimental data. The former expresses biological knowledge on how interactions of genes are expected to affect the concentration of genes; the prior knowledge captures any known information about specific genes, including interactive potential between two genes and functions of genes, which is normally available from online biological databases. Abducibles are unknown signed-directed regulations between genes (the biological analogy of directed links in the graph example given above). Integrity constraints over the abducibles are of four different categories: (1) constraints that enforce signed-directed regulations to be compatible with existing/established knowledge (e.g. already known regulations or compatibility with known type of regulation of the gene), (2) constraints about compatibility of the signed-directed regulations with experimental data, (3) constraints that express logical consistency of the extracted logical model, and finally category (4) that includes constraints about biological consistency. We describe below each of the components of our ARNI framework.

1.2.2.1. *Prior knowledge*

Gene interactions can be of two types, protein-DNA interactions (PDI) and protein-protein interactions (PPI). PDI are directed links from a transcription factor to a regulated gene, whereas PPI interactions are undirected links between proteins. Signed-directed regulations between genes can be of two types, *compatible* and *competitive*. These types of gene regulations are in general unknown and therefore constitute the incomplete part of prior biological knowledge. Computing a regulatory network that conforms with observed gene expression data means discovering those

unknown signed-directed regulations between genes, or signed-directed links, that cause the observed data, in a way that is consistent with given biological constraints.

The domain of genes considered in our abductive task is given by the set of genes that are present in a biological experiment. We denote this set with $\mathcal{G}$. Known potential interactions between genes are expressed in the prior knowledge as logical facts of the form $\texttt{interactive\_potential}(\texttt{g}_\texttt{i}, \texttt{g}_\texttt{j})$, which state that "there is a form of interaction between genes $\texttt{g}_\texttt{i}$ and $\texttt{g}_\texttt{j}$". PDI interactions are normally unidirectional whereas PPI interactions are bidirectional. Therefore our prior knowledge will include only one ground fact of the form $\texttt{interactive\_potential}(\texttt{g}_\texttt{i}, \texttt{g}_\texttt{j})$ for any known potential PDI interaction, and for any known PPI interaction between pairs of genes $\texttt{g}_\texttt{i}$ and $\texttt{g}_\texttt{j}$, two ground facts $\texttt{interactive\_potential}(\texttt{g}_\texttt{i}, \texttt{g}_\texttt{j})$ and $\texttt{interactive\_potential}(\texttt{g}_\texttt{j}, \texttt{g}_\texttt{i})$. We denote with IP_{prior} the following set of ground facts:

$$IP_{prior} = \{\texttt{interactive\_potential}(\texttt{g}_\texttt{i}, \texttt{g}_\texttt{j}) \mid \texttt{g}_\texttt{i} \in \mathcal{G} \text{ and } \texttt{g}_\texttt{j} \in \mathcal{G}\} \quad [1.1]$$

It is important to note that the information of `interactive_potential` in the prior knowledge does not fully capture the regulatory effects between genes as it does not express the type of signed-directed interaction between two genes. This information is expressed by our abducibles, and it has to be consistent with any known information about the regulatory potential of a gene. Known regulatory potential of a gene is extracted from online biological databases and expressed in our prior knowledge as ground facts of the form $\texttt{regulatory\_potential}(\texttt{g}_\texttt{i}, \texttt{s})$ where $\texttt{g}_\texttt{i}$ is a gene and $\texttt{s}$ is the type of regulation, which can be 1 (for *activation*) or -1 (for *inhibition*). For instance, the statement $\texttt{regulatory\_potential}(\texttt{g}_\texttt{i}, 1)$ (respectively, $\texttt{regulatory\_potential}(\texttt{g}_\texttt{i}, -1)$) in the prior knowledge captures the fact that the effect of the regulator gene $\texttt{g}_\texttt{i}$ on any other gene can only be of type activation (respectively, inhibition). When no information about the regulatory potential of a gene is included in the prior knowledge (because unavailable), then that gene can be assumed to have either positive or negative effect on any other gene. Again, our abductive inference process takes into account these two possibilities when reasoning about the effects of gene interactions and, as explained later in section 1.2.2.3, integrity constraints will guarantee that such assumptions are made in a consistent manner. We denote with RP_{prior} the following set of ground facts:

$$RP_{prior} = \{\texttt{regulatory\_potential}(\texttt{g}_\texttt{i}, \texttt{s}) \mid \texttt{g}_\texttt{i} \in \mathcal{G} \text{ and } \texttt{s} \in \{-1, 1\}\} \quad [1.2]$$

As mentioned above signed-directed regulations between genes are the unknown abducibles. It is possible, however that for some pair of genes, say $\texttt{g}_\texttt{i}$ and $\texttt{g}_\texttt{j}$ in $\mathcal{G}$, specific information exists about their signed-directed regulation. Any such knowledge is expressed as atoms of the form $\texttt{established\_regulation}(\texttt{g}_\texttt{i}, \texttt{g}_\texttt{j}, \texttt{s})$ where $\texttt{g}_\texttt{i}$ and $\texttt{g}_\texttt{j}$ are different genes in $\mathcal{G}$ and $\texttt{s}$ is again the type of regulation. For

instance, a ground atom of the form $\texttt{established\_regulation}(\texttt{g}_\texttt{i}, \texttt{g}_\texttt{j}, 1)$ states that $\texttt{g}_\texttt{j}$ is a known activator of $\texttt{g}_\texttt{i}$, whereas a ground atom of the form $\texttt{established\_regulation}(\texttt{g}_\texttt{i}, \texttt{g}_\texttt{j}, -1)$ denotes that $\texttt{g}_\texttt{j}$ is a known inhibitor of $\texttt{g}_\texttt{i}$. Again, our integrity constraints guarantee that abduced signed-directed regulations between genes are consistent with any already known type of regulation. We denote with ER_{prior} the following set of ground facts:

$$ER_{prior} = \{\texttt{established\_regulation}(\texttt{g}_\texttt{i}, \texttt{g}_\texttt{j}, \texttt{s}) \mid \texttt{g}_\texttt{i} \in \mathcal{G}, \texttt{g}_\texttt{i} \in \mathcal{G} \text{ and } \texttt{s} \in \{-1, 1\}\} \quad [1.3]$$

Finally, information about experimental data is also part of the prior knowledge. This includes the expression value of the genes measured in an experiment[2], represented using ground facts of the form $\texttt{exp\_data}(\texttt{g}_\texttt{i}, \texttt{s})$, where $\texttt{g}_\texttt{i}$ is a gene and $\texttt{s}$ is the *state* of the gene, which can be equal to 1 (respectively, -1) to denote that the expression value of $\texttt{g}_\texttt{i}$ has increased (respectively, decreased). Specific information about genes that have been potentially *overpowered* during the biological regulation process is also computed from the experimental data and added to the prior knowledge as ground facts of the form $\texttt{overpowered}(\texttt{g}, \texttt{g}_\texttt{i}, \texttt{g}_\texttt{j})$, where $\texttt{g}$, $\texttt{g}_\texttt{i}$ and $\texttt{g}_\texttt{j}$ are different genes. This fact captures the biological notion that the effect of gene $\texttt{g}_\texttt{i}$ on $\texttt{g}$ has overpowered the effect of gene $\texttt{g}_\texttt{j}$ on $\texttt{g}$. For this, to occur the degree of interdependency between the expression value of $\texttt{g}_\texttt{i}$ and $\texttt{g}$, multiplied by the degree by which the expression value of gene $\texttt{g}_\texttt{i}$ has increased, is higher then the inter-dependency between the expression value of $\texttt{g}_\texttt{j}$ and $\texttt{g}$, multiplied by the degree by which the expression value of gene $\texttt{g}_\texttt{j}$ has decreased. This function is computed using statistical packages provided by R/Bioconductor project [GEN 04]. Last, but not least, experimental data also includes the notion of a subset of genes, within the large pool $\mathcal{G}$, that are considered to be *seed* genes. This information is represented using ground facts of the form $\texttt{seed}(\texttt{g}_\texttt{i})$, which states that gene $\texttt{g}_\texttt{i}$ is a seed gene.

$$\begin{aligned} ExpData = \{&\texttt{exp\_data}(\texttt{g}_\texttt{i}, \texttt{s}); \mid \texttt{g}_\texttt{i} \in \mathcal{G}, \text{ and } \texttt{s} \in \{-1, 1\}\} \\ \cup \{&\texttt{overpowered}(\texttt{g}, \texttt{g}_\texttt{i}, \texttt{g}_\texttt{j}); \mid \texttt{g} \in \mathcal{G}, \texttt{g}_\texttt{i} \in \mathcal{G} \text{ and } \texttt{g}_\texttt{j} \in \mathcal{G}\} \\ \cup \{&\texttt{seed}(\texttt{g}_\texttt{i}) \mid \texttt{g}_\texttt{i} \in \mathcal{G}\} \end{aligned} \quad [1.4]$$

In summary, the prior knowledge of our ARNI's background knowledge, denoted with B_{Prior}, is given by the union of specific subsets of the sets (1.1)-(1.4).

2 We assume in this paper the logic-based modeling of regulatory network from single experiments. The approach can be easily generalized to cross-experiments problems by extending the formalization of our model with an extra argument to denote the name of the experiment.

1.2.2.2. *Rule-based underlying model*

The core rules of our model seek to connect a set of genes (i.e. the seed genes), which have been affected in a biological experiment, to each other, either directly or indirectly by using the information about PDI and PPI interactions given in the prior knowledge, and to abduce signed-directionality between linked genes that are consistent with the (biological) integrity constraints explained in section 1.2.2.3. This consists of computing all possible paths that connect seed genes within a given maximum length, using the following rule-based logic:

$$\texttt{connect\_seeds}(\texttt{MaxLength}, \texttt{Path}) \leftarrow \texttt{seed}(\texttt{G1}), \texttt{seed}(\texttt{G2}), \\ \texttt{path}([\texttt{G2}], \texttt{G1}, 0, \texttt{MaxLength}, \texttt{Path}) \quad [1.5]$$

$$\texttt{path}([\texttt{H}|\texttt{T}], \texttt{G}, \texttt{CurrLength}, \texttt{MaxLength}, [\texttt{G}, \texttt{H}|\texttt{T}]) \leftarrow \texttt{CurrLength} < \texttt{MaxLength}, \\ \texttt{relevant\_ip}(\texttt{G}, \texttt{H}). \quad [1.6]$$

$$\texttt{path}([\texttt{H}|\texttt{T}], \texttt{G}, \texttt{CurrLength}, \texttt{MaxLength}, [\texttt{NewH}, \texttt{H}|\texttt{T}]) \leftarrow \texttt{CurrLength} < \texttt{MaxLength}, \\ \texttt{relevant\_ip}(\texttt{NewH}, \texttt{H}), \texttt{not seed}(\texttt{NewH}), \quad [1.7] \\ \texttt{member}(\texttt{NewH}, [\texttt{H}|\texttt{T}]).$$

$$\texttt{path}([\texttt{H}|\texttt{T}], \texttt{G}, \texttt{CurrLength}, \texttt{MaxLength}, \texttt{Path}) \leftarrow \texttt{CurrLength} < \texttt{MaxLength}, \\ \texttt{relevant\_ip}(\texttt{G1}, \texttt{H}), \texttt{notseed}(\texttt{G1}), \quad [1.8] \\ \texttt{not member}(\texttt{G1}, [\texttt{H}|\texttt{T}]), \texttt{NewCL is CurrLength} + 1 \\ \texttt{path}([\texttt{G1}, \texttt{H}|\texttt{T}], \texttt{G}, \texttt{NewCL}, \texttt{Path}).$$

Rule [1.5] has the effect of constructing a path within the maximum length boundary (`MaxLength`) that links two seeds genes (i.e. `G1` and `G2`). The path is recursively computed by checking that no gene is revisited more than once (i.e. rule [1.7]), and that only relevant genes, according to the existing prior knowledge of interactive potentials between genes, are added to a path (i.e. rule [1.8]). The latter case is captured by the use of the abducible predicate $\texttt{relevant\_ip}(\texttt{G1}, \texttt{G2})$, and the following integrity constraint:

$$\texttt{ic} \leftarrow \texttt{relevant\_ip}(\texttt{G1}, \texttt{G2}), \texttt{not interactive\_potential}(\texttt{G1}, \texttt{G2}) \quad [1.9]$$

The abducibles $\texttt{relevant\_ip}(\texttt{g}_\texttt{i}, \texttt{g}_\texttt{j})$ identify all the genes from a given pool $\mathcal{G}$ that, according to prior biological knowledge are biologically relevant in regulations that can directly or indirectly affect the given *seed* genes. The use of these abducibles allows us to constrain the space of our regulation network in a biologically meaningful way making the computation process more manageable. Assumptions about $\texttt{relevant\_ip}(\texttt{G1}, \texttt{G2})$ may also be abduced in order to satisfy other

constraints, discussed later, so to guarantee their connectiveness with other genes the following constraint is enforced:

$$\texttt{ic} \leftarrow \texttt{relevant\_ip}(\texttt{G1},\texttt{G2}), \texttt{not in\_path}(\texttt{G1},\texttt{G2}) \quad [1.10]$$

$$\texttt{in\_path}(\texttt{G1},\texttt{G2}) \leftarrow \texttt{connect\_seeds}(\texttt{MaxLength},\texttt{Path}), \texttt{append}(\texttt{Path1},[\texttt{G1},\texttt{G2}|\texttt{Path}])$$

Paths generated by the above clauses are sequences of genes, which are connected with each other according to the abduced $\texttt{relevant\_ip}(\texttt{g}_\texttt{i},\texttt{g}_\texttt{j})$ directed link[3]. But to generate a regulatory network, the directed links have to be signed. The inference of the sign for each abduced directed link is generated by means of the following integrity constraint:

$$\texttt{ic} \leftarrow \texttt{relevant\_ip}(\texttt{G1},\texttt{G2}), \texttt{not signed}(\texttt{G1},\texttt{G2}) \quad [1.11]$$

$$\texttt{signed}(\texttt{G1},\texttt{G2}) \leftarrow \texttt{compatible}(\texttt{G1},\texttt{G2},\texttt{S}), \texttt{sign}(\texttt{S}). \quad [1.12]$$

$$\texttt{signed}(\texttt{G1},\texttt{G2}) \leftarrow \texttt{competitive}(\texttt{G1},\texttt{G2},\texttt{S}), \texttt{sign}(\texttt{S}). \quad [1.13]$$

$$\texttt{ic} \leftarrow \texttt{compatible}(\texttt{G1},\texttt{G2},\texttt{S}), \texttt{not relevant\_ip}(\texttt{G1},\texttt{G2}) \quad [1.14]$$

$$\texttt{ic} \leftarrow \texttt{competitive}(\texttt{G1},\texttt{G2},\texttt{S}), \texttt{not relevant\_ip}(\texttt{G1},\texttt{G2}) \quad [1.15]$$

where predicates $\texttt{compatible}(\texttt{G1},\texttt{G2},\texttt{S})$ and $\texttt{competitive}(\texttt{G1},\texttt{G2},\texttt{S})$ are also abducibles and they fully capture the notion of a signed-directed link between two genes. Note that the above constraints [1.11]–[1.15], together with the constraints on sign consistency given later, define in effect the notion of *relevant interactive potential* between two genes in terms of either compatible or competitive influence. In addition to constraints [1.9]–[1.15], abduced signed-directed links have to be consistent with existing knowledge: for some pairs of genes, the signed-directed link might already be known. In this case, the prior knowledge would include ground instances of the predicate `established_regulation` and any abduced compatible fact will have to be consistent with this prior. This is captured by constraint [1.16]. Similarly, the abduced type of compatible or competitive influence that a gene has on another gene has to be consistent with the type of regulatory potential that that gene is known to have (if any). This is expressed in constraints [1.17]–[1.18]. Constraint [1.19], instead, guarantees that competitive regulations are limited to links with an already known regulatory effect. This is done to further limit the solution space for this abducible. Biologists could remove this constraint whenever they intend to pursue a more exploratory analysis:

$$\texttt{ic} \leftarrow \texttt{compatible}(\texttt{G1},\texttt{G2},\texttt{S}), \texttt{G1}! = \texttt{G2}, \quad [1.16]$$

3 Note that the directionality of the link is expressed by the order of the arguments in the predicate.

$$\texttt{established\_regulation}(\texttt{G1},\texttt{G2},\texttt{S1}),\texttt{S} \neq \texttt{S1}$$

$$\texttt{ic} \leftarrow \texttt{compatible}(\texttt{G1},\texttt{G2},\texttt{S}),\texttt{G1!} = \texttt{G2},\texttt{not regulatory\_potential}(\texttt{G2},\texttt{S}) \quad [1.17]$$

$$\texttt{ic} \leftarrow \texttt{competitive}(\texttt{G},\texttt{G2},\texttt{S}),\texttt{G1!} = \texttt{G2},\texttt{not regulatory\_potential}(\texttt{G2},\texttt{S}) \quad [1.18]$$

$$\texttt{ic} \leftarrow \texttt{competitive}(\texttt{G1},\texttt{G2},\texttt{S}),\texttt{G1!} = \texttt{G2}, \quad [1.19]$$

$$\texttt{not established\_regulation}(\texttt{G1},\texttt{G2},\texttt{S})$$

In summary, the rule-based underlying model, Π of our ARNI approach is given by rules in clauses [1.5]–[1.19]. Constraints in clauses [1.5]–[1.19] are part of the IC component of our abductive problem, of which constraints [1.9] and [1.16]–[1.19] guarantee the compatibility of the abduced signed-directed links with the existing knowledge.

1.2.2.3. *Integrity constraints*

As mentioned at the beginning of section 1.2.2.1, our abductive problem is to identify unknown compatible and competitive gene regulations (i.e. signed-directed links) that form a regulatory network which consistently satisfies the observed data. The main abducibles in our ARNI approach are therefore ground facts of $\texttt{compatible}(\texttt{g}_\texttt{i},\texttt{g}_\texttt{j},\texttt{s})$ and $\texttt{competitive}(\texttt{g}_\texttt{i},\texttt{g}_\texttt{j},\texttt{s})$, whose first two arguments are genes in $\mathcal{G}$ and the third argument $\texttt{s}$, which is a binary variable over the set $\{1,-1\}$, denotes the causal effect of the interaction between the two genes $\texttt{g}_\texttt{i}$ and $\texttt{g}_\texttt{j}$. For example, an instance of the form $\texttt{compatible}(\texttt{g}_1,\texttt{g}_2,1)$ (respectively, $\texttt{compatible}(\texttt{g}_1,\texttt{g}_2,-1)$) means that gene $\texttt{g}_2$ *activates* (respectively, *inhibits*) gene $\texttt{g}_1$. Abduced sign-directed gene regulations have to be consistent with the four different classes of constraints described in section 1.2.2.2. Constraint of the first class (i.e. compatibility with existing knowledge) are the above constraints [1.9] and [1.12]–[1.16]. We present integrity constraints of classes (b)–(d) and explain their biological relevance.

Activation and inhibition regulations between genes is formalized by instances of the abducibles $\texttt{compatible}(\texttt{G}_1,\texttt{G}_2,1)$ and $\texttt{compatible}(\texttt{G}_1,\texttt{G}_2,-1)$. So, why do we also need to infer competitive influence (i.e. $\texttt{competitive}(\texttt{G}_\texttt{i},\texttt{G}_\texttt{j},\texttt{S})$)? The biological motivation for modeling competitive gene influences is to reflect the underlying structure of real biological networks, where crosstalk between signaling pathways, regulatory feedback mechanisms and redundancy are common aspects of a biological system. The incoherent network motifs of feed forward loop (FFL) and negative feedback loop, discussed in section 1.1, inherently consist of competitive gene influences. Any inference method aiming to detect such motifs, needs to either rely on multiple experiments to expose each of the influences individually, or to model the concept of competitive gene influences explicitly, as done in our approach.

The latter case has the added advantage that network motifs can be detected using less experimental data, and competitive gene influences can be placed within the same network as their compatible counterparts. Including these regulations in the final solutions is also important for the applicability of the inferred networks within the scope of planning future experiments and network based drug discovery/repositioning. Following an experimental perturbation, competitive gene influences could compensate for the intended experimental response and thus rendering the experiments non-informative. Similarly, competitive gene influences that are enhanced in the presence of a drug might lead to unforeseen side effects. Overlooking the problem of competitive gene influences can result in inconsistencies between the observed and predicted drug effects/experiment outcomes, hindering the process of knowledge discovery.

The inference of $\texttt{compatible}(\texttt{G}_\texttt{i}, \texttt{G}_\texttt{j}, \texttt{S})$ and $\texttt{competitive}(\texttt{G}_1, \texttt{G}_2, \texttt{S})$ has to comply not only with existing knowledge, but also with experimental data, biological principles of *sign consistency* and internal logical consistency of the model. These principles are expressed in our ARNI approach as domain specific *integrity constraints*. This is where our ARNI approach benefits from its abductive logic-based inference process. According to the type of biological experiments and investigation in hand, different classes of constraints could be added or deleted without affecting the formal framework (e.g. in order to compute specific types of regulatory networks (e.g. networks with specific regulatory motifs: "and" gates, "or" gates, etc.).

One of the key biological principles is *sign consistency*. Sign consistency states that inferred gene interactions must satisfy two main gene dependency rules: *compatible gene influence* and *competitive gene influence*. The compatible gene influence postulates that the state of a target gene G1 is directly related to the state of an activator G2 and inversely related to the state of an inhibitor G2. To specify these principles we make use of an additional predicate, called state, which takes two arguments, a gene and a state value. The state value of a gene can be 1 to signify the gene expression is increased, and value -1 to represent that the gene expression has decreased. A ground literal of the form $\texttt{state}(\texttt{g}_1, 1)$ means that the expressive value of gene $\texttt{g}_1$ has increased during the experiment. Since not all states of relevant genes are measurable in an experiment, the information about the state of each gene in our pool is only partially present in our background knowledge. To guarantee full consistency of our regulatory network, the state predicate is therefore considered to be an additional abducible. Integrity constraints for sign consistency include:

$$\texttt{ic} \leftarrow \texttt{compatible}(\texttt{G1}, \texttt{G2}, 1), \texttt{state}(\texttt{G1}, \texttt{S1}), \texttt{state}(\texttt{G2}, \texttt{S2}). \texttt{S1} \neq \texttt{S2} \quad [1.20]$$

$$\texttt{ic} \leftarrow \texttt{compatible}(\texttt{G1}, \texttt{G2}, -1), \texttt{state}(\texttt{G1}, \texttt{S1}), \texttt{state}(\texttt{G2}, \texttt{S1}). \quad [1.21]$$

$$\texttt{ic} \leftarrow \texttt{competitive}(\texttt{G1}, \texttt{G2}, 1), \texttt{state}(\texttt{G1}, \texttt{S1}), \texttt{state}(\texttt{G2}, \texttt{S1}). \quad [1.22]$$

$$\texttt{ic} \leftarrow \texttt{competitive}(\texttt{G1}, \texttt{G2}, -1), \texttt{state}(\texttt{G1}, \texttt{S1}), \texttt{state}(\texttt{G2}, \texttt{S2}), \texttt{S1} \neq \texttt{S2}. \quad [1.23]$$

The incompatibility of the competitive gene inference with experimental data implies that the abductive inference of competitive (G1,G2,S) cannot be driven by the data. The search space explosion in allowing the competitive regulators to be abduced without any constraints is practically prohibitive and hinders the usability of the inferred networks. Therefore, the following constraints and related definitions [1.24]–[1.28] are included in our model to capture two typical cases of competitive regulators that bypass the sign consistency principle:

$$\texttt{ic} \leftarrow \texttt{competitive}(\texttt{G1},\texttt{G2},\texttt{S}), \texttt{not op\_exception}(\texttt{G1},\texttt{G2},\texttt{S}) \quad [1.24]$$

$$\texttt{op\_exception}(\texttt{G1},\texttt{G2},\texttt{S}) \leftarrow \texttt{overpowered}(\texttt{G1},\texttt{G2},\texttt{G3}), \texttt{compatible}(\texttt{G1},\texttt{G3},\texttt{W}) \quad [1.25]$$

$$\texttt{op\_exception}(\texttt{T},\texttt{R},\texttt{S1}) \leftarrow \texttt{compatible}(\texttt{T},\texttt{R2},\texttt{S3}), \texttt{compatible}(\texttt{R2},\texttt{R},\texttt{S2}), \quad [1.26]$$
$$\texttt{iff}(\texttt{S1},\texttt{S2},\texttt{S3})$$

$$\texttt{op\_exception}(\texttt{T},\texttt{R2},\texttt{S3}) \leftarrow \texttt{compatible}(\texttt{T},\texttt{R},\texttt{S1}), \texttt{compatible}(\texttt{R2},\texttt{R},\texttt{S2}), \quad [1.27]$$
$$\texttt{iff}(\texttt{S1},\texttt{S2},\texttt{S3})$$

$$\texttt{op\_exception}(\texttt{R2},\texttt{R},\texttt{S2}) \leftarrow \texttt{compatible}(\texttt{T},\texttt{R},\texttt{S1}), \texttt{compatible}(\texttt{T},\texttt{R2},\texttt{S3}), \quad [1.28]$$
$$\texttt{iff}(\texttt{S1},\texttt{S2},\texttt{S3})$$

Integrity constraint [1.24] guarantees that competitive regulators are only inferred if there is an exception that holds. A gene, say g_1, can have an inconsistent state with respect to the state of its regulator, say g_2 provided that there exist at least one other compatible gene, say g_3 that consistently regulates g_1, hence overpowering the influence of g_2. This principle is captured by rule [1.25]. Exceptions of the above form, are derived from the data by means of an overpowered influence function that determines the truth of the condition $\texttt{overpowered}(g, g_i, g_j)$. Once pre-calculated from the data (see section 1.2.2.1), this information is added as fact to the prior knowledge to express the biological notion that the effect of gene g_i on g has overpowered the effect of gene g_j on g.

Because of the way the overpowered facts are computed there is the additional implicit constraint that genes that can participate in competitive regulations, must have been observed as either up-regulated or down-regulated. Given the sparsity in the microarray data, where the signal is fragmented due to the noise, and the abstraction of all biological regulation to gene regulation, such situations are not very common. In the absence of additional priors information (e.g. kinetic information, promoter affinities), that can give information on the relative impact of competitive influences, our model includes an additional exception case based on the biological

principle of how competitive regulators can participate in some specified network motifs. These are captured by rules [1.26]–[1.28] that correspond, respectively, to the three scenarios in Figure 1.4, where the dashed links represent the competitive influence link involved in the overpowered exception.

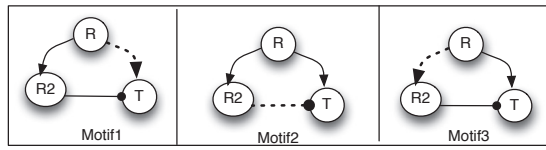

Figure 1.4. *Network motifs of competitive influence*

The sign value of the three sign-directed links that are involved in these motifs have to satisfy one of the predefined incoherent feed forward loop cases, expressed by the fact $\texttt{iff}(\texttt{S1}, \texttt{S2}, \texttt{S3})$ and graphically illustrated in Figure 1.5. Note that the three motif examples given in Figure 1.4 have all the same configuration $\texttt{iff}(1, 1, -1)$. Similar groups of three motifs, one for each of the four possible configurations of incoherent feed forward loops, could happen in regulatory networks.

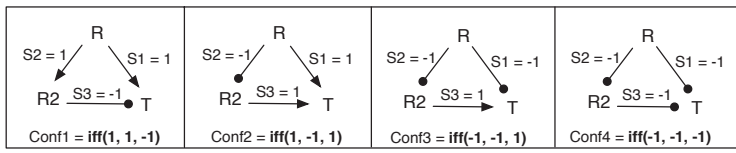

Figure 1.5. *Configurations of incoherent feed forward loops, (iff(S1, S2, S3)*

During the inference process many `compatible` and `competitive` abducibles can be generated. It is important to guarantee that a gene is not assumed to be at the same time a compatible and a competitive regulator of another gene. This is captured by the integrity constraint [1.29]. Similarly, a compatible (respectively, competitive) gene cannot be both activator and inhibitor of another gene. Constraints [1.30]–[1.31] make sure that this principle is satisfied during the inference of signed directed links between genes, whereas constraint [1.32] enforces that a gene can have only one unique state value (i.e. can either decrease or increase its expressive value during a single experiment).

$$\texttt{ic} \leftarrow \texttt{compatible}(\texttt{G1}, \texttt{G2}, \texttt{S}), \texttt{competitive}(\texttt{G1}, \texttt{G2}, \texttt{S}) \quad [1.29]$$

$$\texttt{ic} \leftarrow \texttt{compatible}(\texttt{G1}, \texttt{G2}, 1), \texttt{compatible}(\texttt{G1}, \texttt{G2}, -1) \quad [1.30]$$

$$\texttt{ic} \leftarrow \texttt{competitive}(\texttt{G1}, \texttt{G2}, 1), \texttt{competitive}(\texttt{G1}, \texttt{G2}, -1) \quad [1.31]$$

$$\texttt{ic} \leftarrow \texttt{state}(\texttt{G}, \texttt{S1}), \texttt{state}(\texttt{G}, \texttt{S2}), \texttt{S1} \neq \texttt{S2}. \quad [1.32]$$

The state of a gene is an abducible in our model. This is because given an experiment it is not guaranteed that data about the expression value for each gene will be available (i.e. the background knowledge may include only a subset of the set [1.4]). So for genes that have an expression value the abduced state needs to be consistent with the available experimental data. This is captured by the following constraint [1.33]. For the remaining genes in our identified pool, called in this case *hidden genes*, any of the two states could be abduced provided that the overall set of IC is satisfied.

$$\texttt{ic} \leftarrow \texttt{state}(\texttt{G}, \texttt{S1}), \texttt{exp\_data}(\texttt{G}, \texttt{S2}), \texttt{S1} \neq \texttt{S2}. \quad [1.33]$$

In summary, the full set of integrity constraints included in our ARNI abductive approach, denoted with IC, is given by constraints [1.9]–[1.33].

1.2.2.4. *Inferring signed-directed networks and explanatory reasoning*

As mentioned in section 1.1, in our ARNI approach we can employ abductive reasoning for both inferring a signed-directed regulatory networks from experimental data and enable explanatory scientific reasoning about signal propagations over the generated network in order to help biologists plan the next sets of experiments or improve their understanding of the phenomena in hand. The first abductive reasoning task makes use of the full logical model described in this section. Specifically, it uses as background knowledge the model Π and the knowledge B_{Prior}, which includes a set of experimental data. The set $\mathcal{A}$ of abducibles is the collection of all ground instances of the abducible predicates `compatible`, `competitive` and `state`, together with all ground instances of the auxiliary abducible `relevant_ip`. All these abducible notions are necessary because of the limited available knowledge (i.e. biological information already existing in online databases and the given experimental data), and the desire to generate *realistic* signed-directed regulatory networks that have complex structures (e.g. include feedback loops, competitive regulations, etc.). The set of integrity constraints IC includes all the constraints described in this section. Hence, the question that we are interested in answering in this first type of abductive reasoning task is: *what is a realistic signed-directed regulatory network that has generated the given set of experimental data?* An answer to this question is the abductive inference of a maximal set of signed-directed links between genes with relevant interactive potential that are consistent with the given integrity constraints and the genes' expression level described by the experiment data. The collection of all abduced `compatible` and `competitive` predicates, computed in this answer, formally describe such a signed-directed regulatory network. This abductive reasoning task can be formally defined as follows:

DEFINITION 1.2.– *Abductive inference of regulatory networks. Let the background knowledge $B = B_{Prior} \cup \Pi$, IC be the set of integrity constraints [1.9]–[1.33], $\mathcal{A}$*

be the set of all possible ground instances of the abducible predicates `compatible`*,* `competitive`*,* `state` *and* `relevant_ip`*. An* signed-directed regulatory network inference *is the abductive task* $\langle B, IC, \mathcal{A}\rangle$ *with abductive solution a set* $\Delta \subseteq \mathcal{A}$ *such that:*

$$B \cup \Delta \not\models ic \text{ (}\textbf{consistency}\text{with integrity constraints)}$$

$$B \cup \Delta \cup \{\delta\} \cup IC \models ic \qquad (\textbf{maximality} \text{ of the network})$$

for any $\delta \in \mathcal{A}/\Delta$

The ARNI abductive task may compute more the one possible maximal regulatory network. If the prior knowledge of $\texttt{regulatory\_potential}(\texttt{g}_\texttt{i}, \texttt{g}_\texttt{j})$ is complete for all genes in $\mathcal{G}$ and the gene expression value of every gene is available in the experimental data, then there would be only a single maximal regulatory network that connects all the seeds genes, since, given the constraints, there can be only one possible signed-directed link per pair of genes. But in reality, such complete prior knowledge is not available. More than one maximal network can be generated.

The second abductive reasoning task allows us then to conduct explanatory scientific reasoning about signal propagations over these generated networks. We describe here how this is formally defined. Assuming that a signed-directed regulatory network has been computed, the question that we are interested to answer, in this second abductive reasoning task, is *what are the signal propagations that cause in an existing regulatory network a given collection of seed genes to have certain states?* The answer to this task is a maximal set of states of genes and compatible/competitive signed directed-links in the given network through which signal propagation can occur to cause (i.e explain) a given set of seed genes to be in a given state. This abductive problem assumes a population G to be all the genes that appears in a given regulatory network. The background knowledge B is given by the same logical model considered in definition 1.2, but with the set $Exp\_Data = \{seed(g_i)|$ for some given $g_i\}$ and the signed-directed links that appear in the network, defined as ground instances of $\texttt{established\_regulation}(\texttt{g}_1, \texttt{g}_2, \texttt{s})$ (as they are now part of prior knowledge). The set IC of integrity constraints includes all the constraints [1.9]–[1.33] augmented with the following new constraint [1.34].

$$\texttt{op\_exception}(\texttt{G1}, \texttt{G2}, \texttt{S}) \leftarrow \texttt{compatible}(\texttt{G1}, \texttt{G3}, \texttt{S1}) \qquad [1.34]$$

This constraint captures, together with constraint [1.24], a weaker case of biological consistency for competitive influence than that expressed in constraint [1.25], which does not require experimental data to compute the overpowered function. The set $\mathcal{A}$ of abducibles are in this case given by all ground instances of the predicates `state`, `compatible` and `competitive`.

The additional concept in this abductive reasoning task is the notion of a *query*, i.e. required state of given seed genes. The query is what our abductive reasoning task aims to explain in terms of maximal number of genes' state and sign propagations that cause the given seed genes to reach their given state. For a given query, the Exp_Data will include just the ground facts of the form $\mathtt{seed}(\mathtt{g_i})$ for each gene $\mathtt{g_i}$ in the query. A query can be formally defined as a conjunction of ground atoms of the form $\mathtt{state}(\mathtt{g_i}, \mathtt{s_i})$:

$$Q = \bigwedge_{i=1}^{n} state(g_i, s_i) \qquad [1.35]$$

An abductive reasoning task AC for signal propagation, and abductive explanation for a query Q in AC can be formally defined as follows:

DEFINITION 1.3.– Abductive inference of signal propagations. Let N be a regulatory network. Let Q be a query as defined in [1.35]. Let $B = B_{Prior} \cup \Pi$ but with $Exp\_Data = \{seed(g_i) | g_i$ that appear in Q$\}$ and ER_{Prior} given by the full set of signed-directed links present in N. Let IC be the set of integrity constraints [1.9]–[1.34] and let $\mathcal{A}$ be the set of all possible ground instances of the abducible predicates `state`, `compatible` and `competitive`. An abductive reasoning task for signal propagation is the tuple $\langle B, IC, \mathcal{A} \rangle$. An abductive explanation for Q in AC is any subset Δ of $\mathcal{A}$ such that $M(\Delta)$ is a stable model of $B \cup \Delta$ satisfying the integrity constraints IC and

$$M(\Delta) \models Q \quad \textit{(}\textbf{explanation}\textit{ of the query)}$$
$$M(\Delta) \cup \{\delta\} \models ic \textit{ (}\textbf{maximality}\textit{ of the network)}$$

for any $\delta \in \mathcal{A}/\Delta$

In the special case where the task in definition 1.3 is given as input to a network, N, generated by the inference task in definition 1.2, with respect to the same set of seed genes, the state of the genes inferred through signal propagation will be the same as that considered for the construction of N. In particular, the states of genes that appear in the experimental data used for generating N will be the same as that given in the data. In other words, the signal propagation on the learned network N gives an explanation how the observed experimental data came about. The two abductive tasks, therefore, capture together the notion of abductive inference of a network that explains experimental data.

More generally, definition 1.3 can be used to examine the validity of other biological hypothesis over a (given or constructed) regulatory network. Different queries may be formed allowing the biologists to explain how a given set of gene

states can be caused and through which specific signal propagations; the network itself maybe changed, and changes in the state of the genes may be observed as a result of the addition or elimination of new signed-directed links. Also different domain-specific integrity constraints may be considered in order to analyze signal propagation under different biological principles. Examples of these possibilities are presented in section 1.4 where we illustrate how our ARNI approach can be used to support automated scientific discovery, where the validity of different biological hypotheses can be examined and tested outside the laboratory.

1.3. Evaluation of the ARNI approach

This section describes the results on evaluating different aspects and properties of our ARNI approach. Specifically, in section 1.3.1 we validate the predictive power of our abductive reasoning showing that our approach is capable of extracting the correct regulatory networks from in silico (incomplete) data, in the presence of both biological and experimental noise. The prediction shows a recall of approximately 80% where, instead, existing best network inference methods are capable of predicting approximately 60% of a network from datasets generated with the same *in silico* method. In section 1.3.2 we demonstrate the increased expressive power of our ARNI approach by evaluating its ability to extract regulatory network with a range of complex network motifs structures, which are instead not detectable by other existing inference methods.

1.3.1. *ARNI predictive power*

In this section, we validate the predictive power of our ARNI approach in terms of its robustness against incomplete data and noise in the data. The choice of these two types of validations is due to the fact that, given our logical model, the ability of abductively inferring a gold standard network depends on the ability to retrieve as many of the gold standard links as *relevant* interactive potentials[4] and assign to them the correct signed-directionality, which depends ultimately on the given experimental data. Noise in the data may result in incompatibility between the underlying gene regulation model and the observations and thus affect the assignment of signed-directionality. Incompleteness in the data on the other hand requires special procedures to be able to extend the inferred networks with non-observed genes. The two types of validations are conducted in the following way. We consider a given *gold standard* network. We generate from it different in silico noisy datasets, corresponding respectively to different types of perturbations that could occur in real biological networks, and we evaluate the *recall rate* (R) of the network that ARNI computes from each of these datasets. The recall rate is given by the number of

4 This depends on the extent to which the given seed genes cover the gold standard network.

correct signed-directed links that the abductive reasoning is able to abduce from the dataset with respect the total number of signed-directed links present in the gold standard network. Intuitively, this measure expresses how much of a real underlying regulatory network that the ARNI approach is able to extract from given datasets. We also measure the *relative recall rate* (RR), which gives the percentages of abduced signed-directed links relative to the links that have been identified as relevant in the inferred network.The recall rate is also evaluated with respect to different degrees of sparsity of the dataset. It is shown that although it follows the expected trend of decreasing with the increase of sparsity of the data, even in cases of 80% of missing data, the ARNI system is able to correctly extract 80% of the links present in the gold standard. The same type of experiments are conducted for three different sizes of gold standard networks in order to evaluate how the robustness changes with respect to the size of the network. In each experiment, we assume for simplicity that the prior knowledge is complete for the pool G of genes that appear in the gold standard network[5]. Detailed discussion of the findings are given later in this section.

Datasets and metrics

We have used *in silico* datasets generated using the DREAM project [STO 07]. DREAM is a system biology initiative to provide mechanisms for objective assessment of reverse engineering methods. The DREAM project defines annual challenges consisting of a set of in silico networks, with realistic network structure, of varied size and complexity and the corresponding simulated experimental data [MAR 09] .

In our experiments, we have considered 11 different network topologies taken from the DREAM 3 [PRI 10] and DREAM 4 challenges [MAR 10]. These include five 10-gene networks, three 50-gene networks and three 100-gene networks. Using the GeneNetWeaver simulator provided by the DREAM project [SCH 11], we have generated three different datasets for each of the networks. The *wild-type* dataset contains the (steady-state) gene expression levels in the unperturbed network, and provides the control condition against which all other datasets are compared. The *external-response* dataset contains the (steady state) gene expression levels in the network after all nodes with no incoming links have been activated. The external-response dataset corresponds, for example, to experimental data observed in the case of exposure to environmental factors which, via cell surface receptors, activate intracellular signaling networks. Finally, the *multifactorial* dataset contains the (steady-state) gene expression levels obtained by applying random multifactorial perturbations to the wild-type network. Multifactorial perturbations are simulated by

5 For each link in the gold standard network, an $\texttt{interactive\_potential}(\texttt{g}_\texttt{i}, \texttt{g}_\texttt{j})$ fact, an $\texttt{established\_regulation}(\texttt{g}_\texttt{i}, \texttt{g}_\texttt{j}, -1)$ fact and two facts $\texttt{regulatory\_potential}(\texttt{g}_\texttt{i}, 1)$ and $\texttt{regulatory\_potential}(\texttt{g}_\texttt{i}, -1)$ are added to IP_{prior}, ER_{prior} and RP_{prior}, respectively.

slightly increasing or decreasing the basal activation of all genes of the network simultaneously by different random amounts. This dataset can be thought of as experimental data observed in the presence of multiple perturbations in the network (e.g. multiple drug effects). A total of 50 different multifactorial perturbations were obtained from the 10-gene and 50-gene networks, and a total of 100 different multifactorial perturbations were obtained from the 100-gene networks. Two different types of noise were also considered when generating our in silico datasets: biological and experimental noise. The former was simulated by adding a noise term[6] in the dynamics of the networks. The latter was simulated by adding to the data generated after the simulation a measurement error derived from a noise model similar to that observed in microarrays [SCH 11]. In the case of the *external-response* dataset, we have also considered a deterministic simulation with no added measurement noise.

As per normal practice in real biological experiments, each dataset was simulated three times to obtain replicates required for the application of statistical testing (described below). For each run, the set $ExpData$ of experimental data is given by those genes with an observable change in their state as determined by the statistical testing. All genes with a significant ($p - value < 0.05$) difference between the *wild-type* level and the level observed in an experimental condition (i.e. *external-response* or one of the random perturbations of the *multifactorial* dataset), were considered to be affected by the given perturbation to the network and thus included in $ExpData$. All genes with observed change in their state were also specified as seed genes.

The recall R and related recall RR were defined as follows. Let N be the number of signed-directed links present in the gold standard network. For each experiment run we have calculated the TP (i.e. true positive) number of links, which occur in the *gold standard* network and have been abduced with correct sign and direction, and the RIP number of links, which occur in the *gold standard* network, and have been abduced as `relevant_ip` in Δ. Given these values, the *recall* is defined as $(TP \times 100)/N$ and the relative recall as $(TP \times 100)/RIP$.

1.3.1.1. *Prediction under biological and experimental noise*

In this set of experiments we have validated the ability of our approach to infer networks in a noisy experimental setup. We have used the *external-response* datasets described above. Figure 1.6 reports the recall rates and relative recall rates for the eleven networks considered and compares the results in the cases of no added noise and experimental noise. The table shows that the relative recall rate is consistently higher than 90%, reflecting the fact that the majority of signed-directed links retrieved as relevant have a correct sign propagation causal effect. That is, the expected state of the genes and the experimental observations correspond, validating

6 We used a coefficient of noise term in stochastic differentials equations equal to 0.05.

the sign consistency principle as an appropriate model for gene regulation in the presence of biological and experimental noise. As the size of the network increases it becomes progressively more difficult to infer the *gold standard* from the data derived from a single experiment, namely the recall rate decreases. But given the high relative recall rates, this decrease is primarily due to the inability of the seed genes to cover the whole network. Normally, biologists have at hand data for more than one experimental condition and thus in large scale networks the low coverage of the seed genes can be compensated with the higher number of experiments. In fact, the DREAM challenges request the inference of networks using a complete set of multifactorial experiments.

	10 Node Networks					50 Node Networks			100 Node Networks		
	Net1	**Net2**	**Net3**	**Net4**	**Net5**	**Net6**	**Net7**	**Net8**	**Net9**	**Net10**	**Net11**
Size	15	16	15	13	12	77	160	173	176	249	195
No Noise											
Recall	100	87.5	93.33	100	100	67.53	70	69.94	82.39	34.94	52.31
Relative Recall	100	93.33	93.33	100	100	98.11	91.06	93.8	91.77	89.69	86.44
Noisy Data											
Recall	100	68.75	93.33	100	100	54.55	67.5	57.8	71.59	29.72	39.49
Relative Recall	100	100	93.33	100	100	97.67	90	97	96.18	79.56	83.7

Figure 1.6. *Predicting gold standard networks from noisy data*

To validate this hypothesis, we have evaluated, for the multifactorial dataset, the *consensus recall rate*, which is the recall rate across all the experiments for each of the eleven networks[7]. This is defined in a similar way as the recall rate but with the TP parameter calculated from the union of the individual inferred networks across experiments. The results are reported in Figure 1.7.

Figure 1.7 shows that the consensus recall rate has a significant gain over the recall rate from single experiments and outperforms the published DREAM results. The most recent DREAM publication [MAR 12a] suggests that the best inference methods are only capable of predicting approximately 60% of a given gene network, whereas our approach reports a consensus recall rate greater than 85% for all networks considered. Selected 100-node networks can be inferred with a consensus recall rate as high as 97%. To qualify even further the impact that multiple experiments have on the consensus recall rate, we have taken two networks, one of the 50-node and one of 100-node networks. We have measured the gain in the consensus recall rate per additional experiment[8] and plotted in Figure 1.8.

7 It should be noted that consensus recall rate is only applicable for datasets interrogating the same *gold standard* network (e.g. Net6 from the 50-node networks.

8 Note that the selection of the added experimental data has been performed randomly.

		10 Node Networks					50 Node Networks			100 Node Networks		
		Net1	Net2	Net3	Net4	Net5	Net6	Net7	Net8	Net9	Net10	Net11
Recall range of single expts.	Min:	0.00	0.00	0.00	0.00	0.00	1.29	8.75	1.73	11.93	13.65	29.74
	Max:	66.66	93.75	86.66	69.23	100.00	53.00	56.87	63.58	40.91	44.18	55.90
Concensus Recall using all expts.		93.3	100	100	100	100	85.71	92.5	96.53	88.7	91.16	96.92

Figure 1.7. *Recall rate for the complete set of multifactorial experiments*

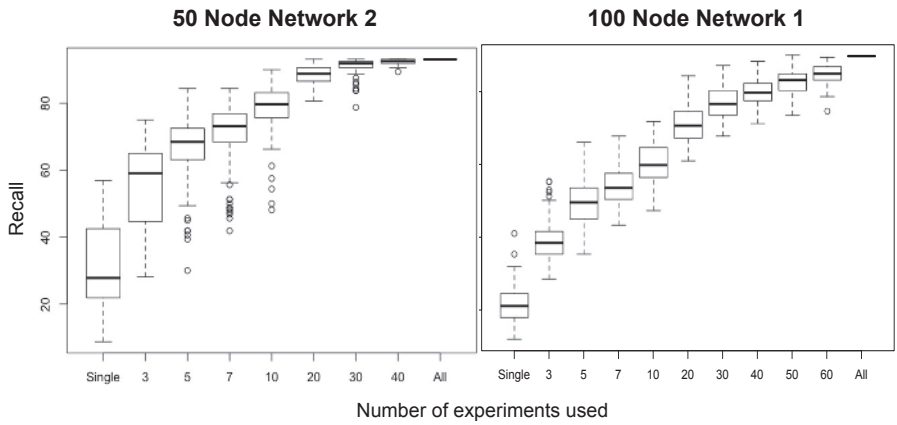

Figure 1.8. *Recall rate relative to the number of individual experiments used*

Figure 1.8 shows that the gain in the consensus recall rate follows a nonlinear trend with more significant gain in the smaller network versus the smaller incremental (but still nonlinear) in the larger network. Increasing the sample size from a single experiment to three experiments (i.e. $n = 3$), results in a 2-fold increase in the consensus recall rate; a further increase in sample size to $n = 5$ results in a 1.16 fold change increase in recall. This is a highly desirable property given the costs in running biological experiments. For example, we see that with 20% of the experiments, we can get between 70 to 80% recall. As discussed above, with further increases in sample size the prediction rate reaches a plateau, which is much higher than that of other approaches.

1.3.1.2. *Prediction under incomplete data*

For a given network, a single experimental dataset normally reveals only a fragmented view of the underlying gene network. The sparsity of the experimental data is partly due to regulation at the protein level which is not captured in gene expression data. Statistical (thresholding) and experimental noise, as well as the inherent robustness and redundancy properties of biological networks that may mask changes in gene effects [ROT 13], also contributes to the sparsity of data. An inference method should be robust to missing non-observed genes and be able to

retrieve biologically consistent information about their states. We validate here the ability of our ARNI approach to infer state information about genes that are not included in the given $ExpData$ set. We have done so by pooling all individual experiments in the *multifactorial* datasets creating a sample size of 700 points (50 experiments for each of the 10-gene and 50-gene networks and 100 experiments for each of the 100-node networks). For each experiment, the percentage of non-observed data is given by the percentage of non-seed genes over the total number of genes in the *gold standard* network. We have run ARNI and computed the recall rate.

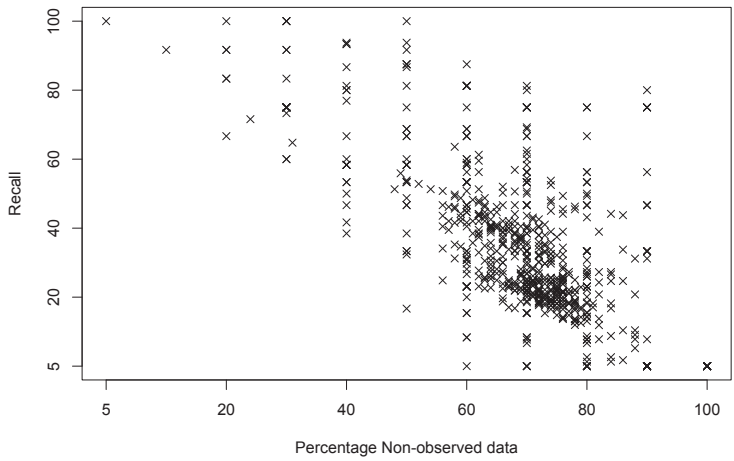

Figure 1.9. *Recall rate under decreasing number of experimental observed genes*

Figure 1.9 shows a general trend of decreasing recall rate with the increase of percentage of missing data. For some percentage of non-observed data, the recall rate varies quite a lot. This reflects that our approach is sensitive to the specific choice of seed genes. The random nature of the perturbation effects in the *multifactorial* datasets leads to some datasets performing particularly badly (i.e. the random effects chosen were inconsistent with the underlying topology of the network). The robustness of our approach to missing data depends on the topological location and the distribution of the seed genes. With an appropriate choice of seed genes recall rates can be in the range of 80% for datasets with as high as 80% missing data. The desired properties of seed genes is to be widely distributed across the networks and to include both upstream and downstream genes. In these experiments the choice of seed genes was not controlled. They were determined by the experimental noise in the data and the biological perturbations in order to be as close as possible to realistic scenarios in real world applications. So, we can only speculate that, under this

scenario, the combination of depth bounded paths between seed genes and the declarative nature of our logic representation would allow the inference of a large proportion of non-observed genes.

In summary, we have shown that our approach is robust to noisy and incomplete data, and it can achieve higher recall rate than established techniques while requiring less experimental data.

1.3.2. *ARNI expressive power*

In section 1.1, we stated that the high-level declarative model used in our ARNI approach overcomes the limitations of existing methods in two ways: it is *more expressive*, in the sense that it enables the inference of networks with more complex regulatory structures, and it is *modular*, as it allows the logic model to be easily adapted to new available information (e.g. addition of new constraints). In this section we substantiate these claims by a series of experiments that demonstrate the expressiveness and modularity of our approach.

1.3.2.1. *Network motif representations*

Network properties and dynamics are determined by recurrent patterns of interactions known as *network motifs*. In section 1.1, we have argued that an effective method of inference has to be able to extract from experimental data regulatory networks that incorporate such motifs. Network motifs can be of different structural complexity. In this chapter we consider motifs given by 3-node feedback loops (FL) and 3-node feed-forward loops (FFL), which contain cascades, fan-in and fan-out components[9].

A regulatory network is said to exhibit a given motif type (or the motif type occurs in the network) when all the signed-directed links that comprise the motif appear in the network. Similarly, a regulatory network inferred by our ARNI approach is said to exhibit a given motif type (or the motif type occurs in the inferred network) when all the signed-directed links that comprise the motif are included in the abduced `compatible` and `competitive` signed-directed links. So, given a gold standard network with n occurrences of a network motif, the ability of ARNI to detect a given motif is measured by the notion of *motif detection rate*. This is the percentage of occurrences of a motif in the inferred network with respect to the number of occurrences of the same motif in the gold standard network. As `compatible` and `competitive` links can only be abduced within the scope of the abduced `relative_ip` that connect seed genes, we have also considered an additional measure, called *motif inclusion rate*. This is the percentage of occurrences of a given motif within the `relevant_ip` of an inferred network with respect to the

9 These are among the most complex motif structures.

number of occurrences of the same motif in the gold standard network[10]. Given these two measures a *normalized detection rate* has also been computed as the ratio between the motif detection rate and the motif inclusion rate. We have considered the gold standard networks described in section 1.3.1 and verified that with respect to the four types of motifs described in Figure 1.10 the normalized detection rate of our approach is above 75%.

	Positive FL	Negative FL	Coherent FFL	Incohernet FFL
				
	n=9	n=4	n=267	n=342
Multifactorial consensus network				
Inclusion	100	100	98.88	98.54
Detection	100	100	95.5	96.78
Normalised Detection	100	100	96.6	98.21
External response networks				
Inclusion	88.09	100	52.81	49.41
Detection	88.89	75	46.06	48.53
Normalised Detection	100	75	87.23	98.22

Figure 1.10. *ARNI's network motif detection profile*

For each of the four motifs shown in Figure 1.10, we have calculated the number of its occurrences across all the *gold standard* networks. A total of nine positive feedback loops, four negative feedback loops, 267 coherent feed-forward-loops and 342 incoherent feed-forward loops were identified. Using the networks inferred in the *multifactorial* and *external-response* experiments, we have calculated the three parameters *motif inclusion rate*, *motif detection rate* and *normalized detection rate* per each motif type. Results are shown in Figure 1.10.

The nearly perfect detection rates observed for the consensus networks indicates that ARNI has no built-in assumptions that forbid the detection of any of the motif types considered, thus validating the hypothesis that ARNI can detect a range of complex network structures known to be present in biological networks. The only lower values are for the coherent FFL and incoherent FFL motifs. In these cases the network for the external response experiment has a detection rate below 50% but it still has high value of the normalized detection rate. This indicates that the low detection rate is due to the inability of the given seed genes to cover the whole network, instead of failure of the constraints for biological consistency, [1.20]–[1.23] and [1.24]–[1.28]. We have verified this hypothesis by rerunning, for the incoherent

10 Note that because of the biological consistency constraints, these two measures do not necessarily give the same results.

FFL motif, the same experiments but without constraints [1.26]–[1.28]. We have found a marked reduction in the detection rate but the same inclusion rate (inclusion= 49,41%, detection=5%). This is because, as shown in Figure 1.11, when these constraints are not considered the ability to infer `competitive` influence is much reduced (i.e. both the median and maximum recall rate of competitive signed-directed links is lower when [1.26]–[1.28] are not included in the model). This not only highlights the importance of modeling competitive gene influences in general, as it increases the recall rate, but it also shows the relevance of competitive influence for the detection of network motifs.

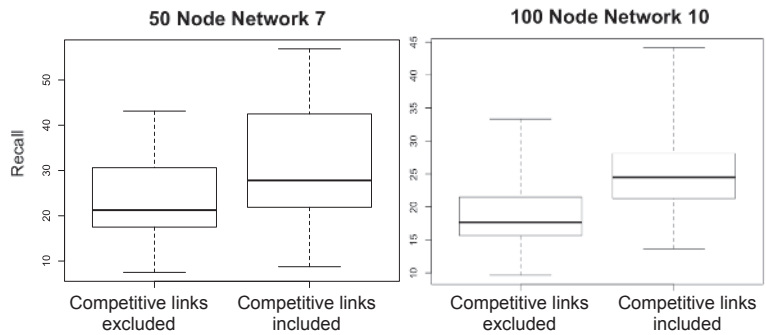

Figure 1.11. *Effect of reasoning over competitive gene influenceson recall rate*

As is the case for the recall rate, the motif detection rate depends on the set of seed genes given to the abductive inference task, as different choices of seed genes may lead to different proportions of missed signed-directed links. To exclude the possibility of a systematic error in the evaluation of the detection rate, we have also tested for over-representation of motif links within the missed signed-directed links in order to answer the following question: *is a motif edge more likely to be missed over a non-motif edge?* Absence of systematic errors would be indicated by equal probability of missing a motif and a non-motif link.

All links in the *gold standard* networks were classified as motif link or non-motif link, depending whether or not they occurred in any of the four tested motifs. Then, the two worst performing (i.e. lower recall rate) *multifactorial* datasets (50-node Net-6 and 100-node Net-9 in Figure 1.7), were considered, and for each random perturbation experiment with respect to these two datasets, we labelled the abduced signed-directed links that also appear in the corresponding gold standard as *inferred* and the others as *non-inferred*. We then performed a chi-squared test on the two factors (motif link and inferred link) to test if there is an over-representation of a particular motif in the links not detected (i.e. false negative). Specifically, we tested whether the ratio of inferred over non-inferred links for motif edges was lower than the ratio of inferred over non-inferred links for non-motif edges. Out of the 150 experiments tested, none of the

p-values was significant (i.e. p-value < 0.05). Hence, we have been able to conclude that our validation of ability to predict motifs was not affected by the particular signed-directed links that were missed.

These experimental results are also particularly promising in underlying the advantage of our logic-based inference approach versus other existing inference methods for regulatory networks. Results published in [MAR 10, MAR 12a] on similar network motif analysis for existing methods have demonstrated that these methods do indeed suffer from systematic errors in detecting feed-forward loops, cascades (i.e. incorrect prediction of shortcuts) and fan-in motifs (i.e. missed regulation between two genes). In our approach the use of prior knowledge on interactive potentials has helped in overcoming this shortfall. To substantiate even further the improvement that our logic-based approach provides in detecting network motifs versus existing approaches, we have compared our results with respect to those achieved using a benchmarking method in physical network inference, referred to as PNM [YEA 04]. In this case we have chosen a network, active in T-cells, that controls T-cell differentiation into two different subtypes. The T-cell network topology is illustrated in Figure 1.12(a). The network includes multiple feedback loops (positive and negative), feed-forward loops and cascades that lead to the regulation of a gene either via transcriptional regulation (triangles) or post-translational regulation (circles). We have applied our ARNI approach and PNM method on simulated data (under no noise conditions) with STAT4 and the three sources, INFγ, IL4 and IL12, in diamond shape[11] as seed genes. ARNI was able to infer the entire gold standard network (Figure 1.12(b)) whereas the PNM approach was able to infer only a partial network (Figure 1.12(c)).

The missed and mislabeled links in Figure 1.12(c) can be attributed to specific limitations of the PNM modeling. The PNM approach can only infer simple paths, so it does not support feedback loop detection. In fact, five of the eight missed links are part of a feedback loop. In addition, the PNM modeling rules impose the restriction that the last link in a path from source to target should be a transcription factor (triangle). But in the gold standard, the IL12R regulation of STAT4 is at the post-translational level causing the path linking IL12 to STAT4 to be missed in the PNM output network. This is clearly a limitation of the PNM, as post-translational regulation has been shown to be an important component in integrated networks [JOS 10]. Finally, the mislabeled link between IL4 and IL4R demonstrates the inability of PNM to reason about competitive gene influences. IL4R is under the competitive regulation of IL4 and SOCS1, and in the particular dataset, SOCS1 overpowers IL4 to determine the required state of IL4R. As PNM does not make use of prior knowledge about regulatory potential, it ends up inferring sign consistencies

11 The background knowledge in each case included complete knowledge of interactive potential and regulatory potential with the expressive level as depicted in the figures (i.e. dark grey for up regulated and light grey for down regulated).

that are against the known IL4's regulatory potential. ARNI has correctly inferred the link between IL4 and IL4R as overpowered activation.

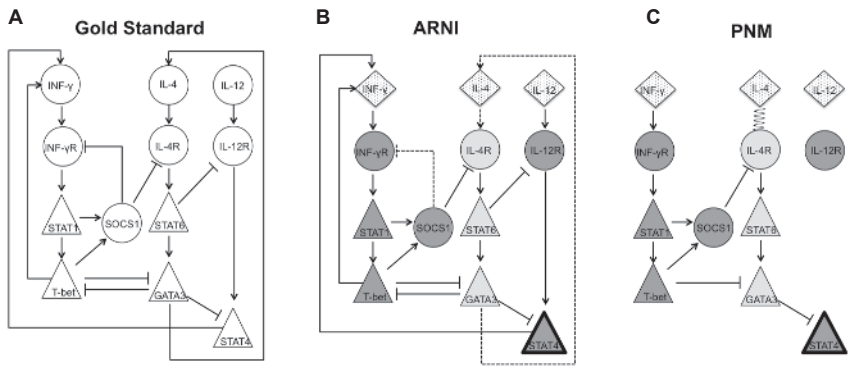

Figure 1.12. *Inference of T-cell differentiation network a) using ARNI b) and PNM [YEA 04] c). The query gene is shown with black border. Diamonds denote source of perturbation, eclipses denote proteins, triangles denote transcription factors. Arrow types denote the regulatory effect: regular (activator) and cut (inhibitor). Dashed lines denote overpowered influences. Wavy lines denote incorrectly inferred link*

In summary, we have shown that ARNI can infer complex regulatory structures, achieving improved expressiveness over existing methods. The non-restrictive nature of the symbolic representation, coupled with reasoning over competitive gene influences and prior knowledge are key features of our approach for the detection of network motifs.

1.3.2.2. *Representing complex interactions*

As it transpires from section 1.2, the constraints of our logical model are grouped by categories of functionally related concepts. Constraints [1.9]–[1.33] form the core of our model and should therefore always be included in any of the two abductive tasks defined in section 1.2. In order to tailor our abductive tasks to specific inferences required by the biologists, additional constraints and assumptions can easily be included in the model without having to redefine it. For instance, in addition to the conventional gene regulation, a biologist might want the inference process to take into account *co-ordinated regulations*. This type of information is typically not available in online biological repositories, and it relies mainly on the knowledge of the biologist. We consider in this section how our model could be extended to allow for two types of coordinated regulations, called, respectively, *allosteric inhibition* and *protein complexes*.

Allosteric inhibition occurs when the binding of one protein on a target prevents the action of another regulator on the same target. Specific instances of allosteric

inhibitions could be easily expressed by constraints of the form given in [1.36] and [1.37]. In this specific case, g_1 is the target gene and g_3 is the binding gene whose activated influence stops the regulation of gene g_2 over g_1. In fact, if given the experimental data and the prior knowledge, it is possible to infer that g_3 is up-regulated (i.e. $\texttt{state}(g_3, 1)$ and $\texttt{compatible}(g_3, g_y, s)$ had been consistently abduced for some value of s), then it cannot be consistently inferred that g_2 activates g_1.(i.e. $\texttt{compatible}(g_1, g_2, 1)$ cannot be abduced). Because of the maximization of the abductive solution and constraint [1.36], the inhibition of g_1 by g_3 will be inferred.

$$\texttt{ic} \leftarrow \texttt{compatible}(g_1, g_2, 1), \texttt{compatible}(g_1, g_3, -1) \quad [1.36]$$

$$\texttt{ic} \leftarrow \texttt{state}(g_1, 1), \texttt{state}(g_2, 1), \texttt{compatible}(g_1, g_2, 1), \texttt{not inactivated}(g_3) \quad [1.37]$$

$$\texttt{inactivated}(X) \leftarrow \texttt{state}(X, -1), \texttt{compatible}(X, Y, S)$$

The above constraints essentially enforce a notion of strong inhibition: in the presence of opposite influences to a common target gene, g_1, the activation of g_1 by gene g_2 can only be inferred provided that the inhibition by g_3 cannot be abductively proved. This is captured by constraint [1.37] and the given definition of `inactivated`.

A *protein complex* occurs when two genes bind to each other to form a complex, which then acts on another target gene. The effect of a complex on a target can be of either activation or inhibition. An activating protein complex is only important in explaining the up-regulation of a gene, whereas an inhibitory protein complex is important in explaining the down-regulation of a gene. In situations where one component of an activating (respectively, inhibitory) complex is down-regulated, it is sufficient on its own to explain the down-regulation (respectively, up-regulation) of its target irrespective of the state of the other component in the complex. The behaviordescribed above, can be expressed with constraints [1.38]–[1.39] for activating complex and [1.40]–[1.41] for inhibitory complex. In the case of activating complex, constraints [1.38]–[1.39] ensure that the same type of interaction of both genes forming the complex are inferred (i.e. in this case g_2 and g_3 form a complex and they both have to have the same signed-directed link with g_1). The state of g_1 has to be in this case up-regulated, since the activating effect of a complex is only important for up-regulation. Constraints [1.40]–[1.41] capture the case of inhibiting complex, where the down-regulation (i.e. $\texttt{state}(g_1, -1)$) is instead relevant.

$$\texttt{ic} \leftarrow \texttt{state}(g_1, 1), \texttt{compatible}(g_1, g_2, 1), \texttt{not compatible}(g_1, g_3, 1) \quad [1.38]$$

$$\texttt{ic} \leftarrow \texttt{state}(g_1, 1), \texttt{compatible}(g_1, g_3, 1), \texttt{not compatible}(g_1, g_2, 1) \quad [1.39]$$

$$\texttt{ic} \leftarrow \texttt{state}(g_1, -1), \texttt{compatible}(g_1, g_2, -1), \texttt{not compatible}(g_1, g_3, -1) \quad [1.40]$$

$$\texttt{ic} \leftarrow \texttt{state}(\texttt{g}_1, -1), \texttt{compatible}(\texttt{g}_1, \texttt{g}_3 - 1), \texttt{not compatible}(\texttt{g}_1, \texttt{g}_2, -1) \quad [1.41]$$

Note that constraints [1.38]–[1.41] are conceptually different from constraints [1.36]–[1.37]. The latter enforce the absence of a link, whereas the former enforce the presence of a link. Adding these additional coordinated regulations may result in significant changes in the resulting networks. For instance, taking as an example a protein complex that controls cell cycle, we would need to express that *cyclinE* and *cdk2* form a complex that leads to inactivation of *retinoblastoma* (rb) protein. This can be expressed using constraints [1.42]–[1.43] below.

$$\texttt{ic} \leftarrow \texttt{state}(\texttt{rb}, -1), \texttt{compatible}(\texttt{rb}, \texttt{cyclinE}, -1), \texttt{not compatible}(\texttt{rb}, \texttt{cdk2}, -1) \quad [1.42]$$

$$\texttt{ic} \leftarrow \texttt{state}(\texttt{rb}, -1), \texttt{compatible}(\texttt{rb}, \texttt{cdk2}, -1), \texttt{not compatible}(\texttt{rb}, \texttt{cyclinE}, -1) \quad [1.43]$$

Embedding [1.42]–[1.43] within our logical model for the inference of a bigger network would exclude some scenarios that would otherwise be abductively inferred. In datasets where rb is a non-observable gene, constraints [1.42]–[1.43] would guarantee that only the correct state for rb is inferred, namely rb is inferred as down-regulated in the context of datasets where cdk2 and cyclinE are up-regulated.

In section 1.4 we demonstrate how constraints [1.42] and [1.43] can be used to test specific hypotheses about the cell-cycle pathway.

1.4. ARNI assisted scientific methodology

In this section, we show that our ARNI approach is not only advantageous for network prediction, but also for performing explanatory scientific reasoning about signal propagations and meta-level reasoning over (inferred) regulatory networks. Specifically, in section 1.4.1 we illustrate how ARNI can enhance scientific knowledge discovery. We present examples in which ARNI is used as a "scientific assistant" to help experts rationalize their hypotheses and guide them on identifying further experiments to improve the biological accuracy of the inferred networks. Section 1.4.2 examines how our logical model can also be used to abductively infer discriminating tests to help choose among alternative regulatory networks.

1.4.1. *Testing biological hypotheses*

The ARNI approach provides a general logic-based model of gene regulations that can be applied to any problem of interest. The flexibility and modularity of the

logical framework allow biologists to setup and perform different inference tasks (see definitions 1.2 and 1.3), such as topology inference, state prediction and model checking, within the same framework. We examine, in this section, how our ARNI approach can be used as a scientific assistant in supporting biologists through an iterative, investigative process. Each iteration is composed of the following steps: (1) automated analysis of the correctness of a current network with respect to a desired biological property (i.e. model checking task), (2) human-driven biological assumptions to address any identified counterexample, and automated check that the new assumptions would eliminate the counterexamples and therefore establish the desired biological properties, (3) automated prediction of the effects that the additional biological assumptions would have on the existing network and (4) automated verification of the correctness of the newly inferred network. We illustrate this process by using a regulatory network that is active in T-cells and controls T-cell differentiation into two different subtypes. This is the same network we have referred in section 1.3.2.1 to demonstrate the expressiveness of our approach. The four steps of a single iteration, applied to this example, are illustrated in Figure 1.13. The iteration starts from a given (potentially incomplete) network (Figure 1.13(a) left) and it ends with the correct network (Figure 1.13(a) right).

Before describing the single steps, we give the relevant biological context of this example. T-cells have been shown to exist in a bistable environment and are either in a state of high t-bet phenotype controlled by infγ, or in a state of high gata3 phenotype controlled by il4. Once the network reaches one of these two states, its behaviorbecomes irreversible independently of the stimuli subsequently received [MEN 06]. Because of this bi-stability the abduced regulatory network should not exhibit the property:

high `il4` *leads to low* `t-bet` *expression or high* `inf`γ *leads to low* `gata3`

To verify this bi-stability property, the biologist can start from either of the two possible unstable cases: “high il4 leads to low t-bet expression” or “high infγ leads to low gata3”. Let’s consider the first case. If the regulatory network was correct, this property would not be satisfied even in the presence of high infγ. We have performed this model checking task by using the abductive inference task given in definition 1.3 with respect to Query 1 in Figure 1.13(b). This query is captured by the following logical query and new background knowledge facts:

`Query:`	`New assumptions:`
`state``(il4, 1)` **`state`**`(inf`γ`, 1)` `state(t-bet, -1)`	**`seed`**`(il4)` **`seed`**`(inf`γ`)` `seed(t-bet)`

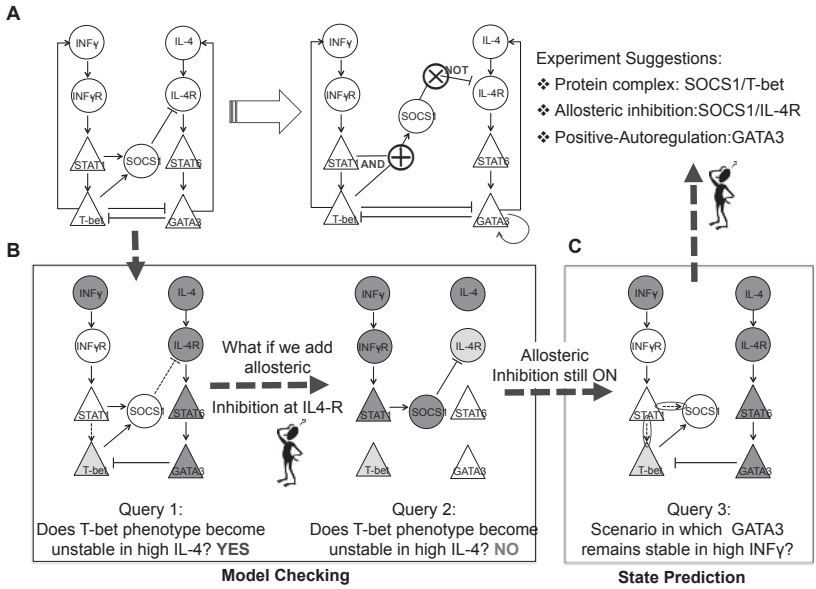

Figure 1.13. *Using ARNI to support the scientific process of extraction of correct regulatory network structures from experimental data. a) A single iteration process, starting from a given (potentially incomplete) network (left) and ending with the correct network(right). b) Left: Automated analysis of the correctness of the current network with respect to the data (i.e. model checking task), Right: Reviewing of counterexamples by biologists can result in possible changes to the network to establish the correctness. c) Automated prediction of the effects that the additional biological assumptions would have on the existing network can result in additional behaviors to be considered in order to guarantee correctness*

The system was able to infer an explanation of how il4 can down-regulate t-bet, which demonstrates lack of biological stability. Such an explanation can therefore be seen as a counter-example to stability. It not only shows that the given topology is insufficient to explain the bi-stability behavior, but it also provides an example behaviorof non-stability, $\texttt{competitive}(\texttt{il4R}, \texttt{socs1}, -1)$. Reviewing this counter-example, the biologist can formulate possible changes to the network. One such hypothesis is that socs1 should act as a strong inhibitor (known in biology terms as allosteric inhibitor) of il4R in order to block il4R signal propagation and establish stability. This additional assumption (i.e. dashed arrow between left and right network in the model checking box of Figure 1.13(b)) can be expressed in our ARNI

model by the addition of the integrity constraint [1.44], which states that "il4 has an effect on il4R, provided that socs1 is not unregulated"[12].

$$\texttt{ic} \leftarrow \texttt{compatible}(\texttt{il4R}, \texttt{il4}, 1), \texttt{activated}(\texttt{socs1}) \qquad [1.44]$$
$$\texttt{activated}(\texttt{socs1}) \leftarrow \texttt{state}(\texttt{socs1}, 1), \texttt{compatible}(\texttt{socs1}, \texttt{X}, \texttt{S})$$

Repeating the same query as above but now with constraint [1.44] ARNI returns no solution, which in logical terms means that the given network with the additional hypothesis can no longer find consistent sign propagations that explain the query "high il4 leads to low t-bet expression". Biologically this is because socs1 can be proven to be upregulated as one of its regulators, stat1, is upregulated. Hence, $\texttt{compatible}(\texttt{il4R}, \texttt{il4}, 1)$ cannot be abduced, removing the only possible path from il4 to T-bet. This second step of model checking proves that a stable t-bet phenotype can be established in the given network under the assumption of allosteric inhibition at il4R, captured by constraint [1.44].

But the initial property of the network is bi-stability. So the added constraint should not affect gata3 stability, which means that property "high ingγ leads to high gata3" should succeed under the same constraint of allosteric inhibition at il4R (or socs1 strong inhibition). The third step of our iterative process is then used to predict the effect that socs1 strong inhibition has on gata3 stability. We have used ARNI to predict sign propagation to explain the query gata3 stability, formalized below, with constraint [1.44] now part of the IC of our abductive task.

Query:	New assumptions:
$\texttt{state}(\texttt{il4}, 1)$ $\texttt{state}(\texttt{inf}\gamma, 1)$	$\texttt{seed}(\texttt{il4})$ $\texttt{seed}(\texttt{inf}\gamma)$
$\texttt{state}(\texttt{gata3}, 1)$	$\texttt{seed}(\texttt{gata3})$ $\texttt{seed}(\texttt{socs1})$
$\texttt{state}(\texttt{socs1}, -1)$	constraint (1.44)

In this case, we seek an explanation that predicts socs1 as downregulated under the assumption that gata3, infγ and il4 are high. The only explanation is that stat1 cannot exert an effect on socs1 on its own, in conjunction with gata3 overpowering stat1's effect on t-bet and thus maintaining t-bet downregulated. The explanation generated by the ARNI system includes the two abducibles $\texttt{competitive}(\texttt{socs1}, \texttt{stat1}, 1)$ and competitive(t-bet, stat1,1), which are depicted with circled arrows in the network in Figure 1.13(c). The biologist may use this prediction to infer a mechanistic biological hypothesis that t-bet and stat1 act on socs1 via a protein complex and that gata3 augments its own expression resulting into stat1 overpowering. These are new biological hypotheses that can be further tested in the lab.

12 This is easy to do because of the possibility of expressing in ARNI complex interactions as discussed in section 1.3.2.2.

1.4.1.1. *Testing cross-talk between signaling pathways*

As a scientific assistant the ARNI approach can be used by biologists to test a variety of biological hypothesis. An interesting biological hypothesis is the *cross-talk* between specific parts of a regulatory network. This type of specific hypothesis is related to the notion of complex regulations, briefly introduced in section 1.3.2.2. The basic idea of a cross-talk hypothesis is the existence of at least a signed-directed link between two genes that belong respectively to two (biological) pathways (i.e. parts of a network) responsible for together regulating a common biological process. Such signed-directed link is needed to either guarantee the combined effect of the biological pathways or to provide alternative activation pathways. Consider, for instance, the biological pathway (or pathway in short) given at Figure 1.14(a). It includes genes p53, waf1, cdk2 and cyclinE, and it regulates the rb gene. We refer to this biological pathway as *cell_cycle*, as it is a well known part of a p53 biological pathway responsible for controlling the cell-cycle and ensuring that cells do not enter into uncontrolled cell proliferation stage. Genes cyclinE and cdk2 form a protein complex that act on the target gene rb. Under normal conditions the *cell_cycle* pathway causes the up-regulation of rb, via the up-regulation of p53 and inhibition of the cyclinE/cdk2 complex.

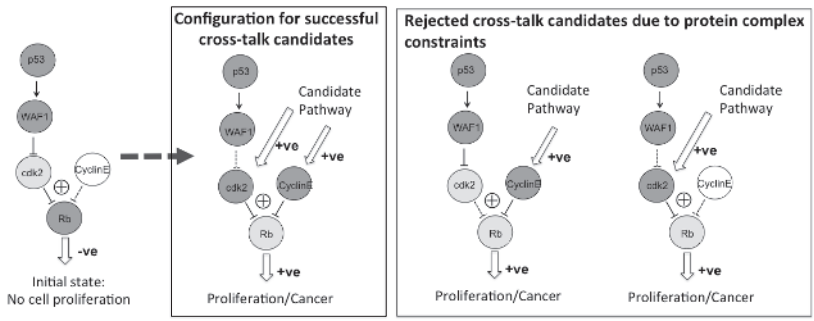

Figure 1.14. *Investigating cross-talk in the p53 cell_cycle pathway. Successful candidates should have a positive effect on both components of the cdk2/cyclinE complex*

Biologists are then interested in identifying the existence of cross-talks between this pathway and another given biological pathway, for instance one regulating genes cyclinE and cdk2, referred here as *candidateBioPathway*, that if it becomes activated is able to reverse the effect of the *cell_cycle* pathway on rb. Identifying cross-talks between the *cell_cycle* and another biological pathway could for instance be significant in revealing potential interventions points for the development of cancer. For instance let's assume that gene p53 is up-regulated. We still would like to guarantee that gene rb is maintained down-regulated. According to just the *cell_cycle* pathway, the up-regulation of p53 would cause the up-regulation of rb, so to guarantee the down-regulation of rb, an appropriate signed-directed link between a

gene in the *candidateBioPathway* and a gene in the *cell_cycle* is desirable (i.e. cross-talk between the two biological pathways). So testing for an explanation of rb being down when p53 is up means identifying a signed consistent cross-talk between the two pathways. The correct situation of cross-talk is illustrated in Figure 1.14(b), where a regulation exist between a gene of the *candidateBioPathway* and both gene cdk2 and gene cyclinE, which are genes of the *cell_cycle* pathway. Note that if protein complex constraints [1.42]–[1.43] were not included the same signed propagation abductive reasoning task would generate solutions where a regulation exists between a gene of the *candidateBioPathway* and only one of the components of the cyclinE/cdk2 complex. These cases are illustrated in Figure 1.14(c).

To use our ARNI approach for this type of hypothesis testing, the following integrity constraint [1.45] can be added to IC, together with the definition, in the background knowledge, of the genes that belong to two biological pathways under consideration. The table below shows an example definition of two biological pathways added to the background knowledge, where *cell_cycle* is abbreviated by ccP, and candidateBioPathway, abbreviated by candP, is assumed to be composed of four genes g1, g2, g3 and g4. We also assume that p53, rb, g1 and g2 are seed genes.

$$\texttt{ic} \leftarrow \texttt{not crosstalk}(\texttt{cell\_cycle}, \texttt{candidateBioPathway}) \qquad [1.45]$$

$$\texttt{crosstalk}(\texttt{P1}, \texttt{P2}) \leftarrow \texttt{inBioPathway}(\texttt{P1}, \texttt{G1}), \texttt{inBioPathway}(\texttt{P2}, \texttt{G2}),$$
$$\texttt{connect\_seeds}(\texttt{MaxLength}, \texttt{Path}), \texttt{member}(\texttt{G1}, \texttt{Path}), \texttt{member}(\texttt{G2}, \texttt{Path})$$

Query:	New assumptions:
state(rb, −1)	seed(rb) seed(p53) state(p53, 1) seed(g1) seed(g2) inBioPathway(ccP, p53) inBioPathway(ccP, waf1) inBioPathway(ccP, cdk2) inBioPathway(ccP, cyclinE) inBioPathway(ccP, rb) inBioPathway(candP, g1) inBioPathway(candP, g2) inBioPathway(candP, g3) inBioPathway(candP, g4)

If our signal propagation, abductive reasoning task finds a solution Δ this will specify a network that connects the given seed genes and that, because of constraint [1.45] has also to include at least two genes of the two given biological pathway. The sign directionality of this network will have to explain the given query of down-regulation of rb when p53 is upregulated. The existence of such a network guarantees the existence of a cross-talk between the two given biological pathways.

In summary, we have presented in section 1.4.1 two key examples of how our ARNI abductive framework can be used as a scientific assistant for testing biological hypotheses. The logical model can be easily adapted to new information that becomes available or relevant to the biological investigation. Biologists can customize the set of seed genes, the set of constraints and/or add hypothetical biological priors in order to explore biological hypotheses by posing queries and

identifying relevant unknown gene influences that consistently explain the given queries. We have illustrated these concepts by considering two examples: the combinatorial regulation of T-cell differentiation network, and the investigation of cross-talk between known biological pathways, where the emphasis has been on the addition of specific domain-dependent integrity constraints. In a similar fashion, the choice of seed genes provides the biologists with a declarative means for controlling the solution space. For instance, explanations for oscillating genes could be inferred via an abductive task that only defines the oscillating genes as seed genes. Our logical-model would in this case be able to infer feedback loops to the oscillating gene, thus revealing novel candidates of negative feedback regulation hidden in experimental data.

1.4.2. *Informative experiments for networks discrimination*

In all the experiments described so far we have assumed complete biological prior knowledge. If such knowledge were incomplete (e.g. `established_regulation` was incomplete), our ARNI abductive task would generate multiple regulatory networks (if any exists) that are consistent with the experimental data and integrity constraints. In real world biological applications, it is indeed often the case that biological knowledge available in online databases is not complete. One of the key challenges is how to decide which of the networks is the "correct" network. In section 1.4 we have presented a process, based on the integration of model checking and state prediction, by which biologists can perform iterative steps of computational investigations to ascertain missing biological assumptions. This process eventually leads to newly discovered information about genes that biologists can further test in the lab. In principle, this process could be applied to the different inferred networks. Clearly, if more networks are inferred during this process, more tests would need to be performed in order to empirically verify the new findings. As biological experiments come with their own costs it is therefore desirable to have a computational mechanism that identifies *key* lab tests. These are tests that, independently of their empirical outcome, can rule out the incorrect inferred networks.

Existing work [MCI 94] have demonstrated that abductive reasoning can be used, in particular in medical diagnosis, for automated test generation. Different classes of tests can be defined in terms of abductive solutions of different specific abductive tasks (see [MCI 94] for further details). Building upon these results, a notion of *discriminating test* can be defined and used for discriminating inferred regulatory networks. In order to rule out a network, among all the inferred alternatives, it is sufficient to disprove just one of its signed-directed links. A test can therefore be seen as a pair of the form $\langle Gene, Observation \rangle$, where $Gene$ is a single gene in the pool $\mathcal{G}$ covered by all the alternative networks, and $Observation$ is a possible state the $Gene$ can be in (i.e. either up-regulated or down-regulated). A discriminating test is therefore a pair $\langle Gene, Observation \rangle$ that is consistent with the current prior

knowledge and experimental data but that, independently of the value of the $Observation$, can refute at least one inferred network. Namely, that for the outcome up-regulated of the tested $Gene$ there would be at least one inferred network N_i that is refuted, and that for the outcome down-regulated there would be at least another inferred network N_j that is refuted. There are several types of biological tests. For example, we can test the state of a gene, by measuring its level in lab, or we could test for causal relationship between genes by performing knock out experiments using $siRNA$. The former is the simplest case and it can be represented as $\texttt{test\_outcome}(\texttt{G}, \texttt{O})$. This fact can be treated as an abducible of a specific abductive task for the inference of discriminating tests, and it can be used to define the state of gene G, i.e. $state(G, O) \leftarrow test\_outcome(G, O)$. The inference of $state(G, O)$ is consistent with some inferred networks but also inconsistent with others. So, given the whole collection of inferred networks, the inference of a discriminating test can be specified as the abductive task of inferring a $\texttt{test\_outcome}(\texttt{G}, \texttt{O})$ that will maximize a given user-defined score priority. The simplest such priority can be in terms of number of inferred networks that the specific test outcome will be inconsistent with (i.e. rules out). A full detailed description of how we can formally define such an abductive task is outside the scope of this chapter, but preliminary investigations have confirmed our intuition that ASP and its optimization mechanisms provide an ideal computational environment for such an abductive task and can be easily built upon the ARNI logical model we have developed.

1.5. Related work and comparison with non-symbolic approaches

A number of statistical approaches to gene network inference exist. These can be divided into three main groups:[13] [LAN 08, SCH 05, MAR 06, FAI 07] use different statistical dependency measures motivated from information theory to infer unsigned-undirected co-expression networks. The works in [FRI 00, FRO 08, WER 08] use probabilistic graphical models to infer joint probability distributions over the observations. Another group of methods use regression analysis to identify best predictors for each gene [IRR 10, KUF 12]. A key advantage of the approach we have presented here, compared with statistical approaches, is the ability to incorporate background knowledge on interactive and regulatory potentials. Having a scaffold of interactions over which unknown networks can be inferred overcomes a number of problems: (1) the sparsity in the input data is addressed and genes in the inferred networks are extended beyond those experimentally measured, thus resulting in more complete networks, (2) the systematic errors in network motifs representation, reported for statistical approaches [MAR 10, MAR 12a], are not present in our method, thus resulting in more realistic networks and (3) the inferred networks are based on physical

13 For a comprehensive review on non-symbolic based approaches, see [ROT 13, HE 09].

molecular interactions and as such are easier to interpret and can reveal the underlying biological mechanism of biological processes.

Abductive reasoning has been suitable for addressing a number of problems in systems biology. [RAY 08, KIN 04] discover the function of genes from auxotrophic growth experiments and synthetic lethal mutations, respectively, [RAY 10, TAM 06], learn/revise metabolic pathways. More relevant to the approach discussed in this chapter, [PAP 05] and [INO 13] use abduction to inferring signaling-transcriptional networks. But existing proposals suffer a high number of false predictions. In [PAP 05], gene dependencies are inferred to explain changes in the gene expression levels using a predefined set of regulators that are allowed to regulate any other gene. No biological prior knowledge is considered during the inference process and the approach does not cater for non-observed genes or concurrent gene regulation. In [INO 13] the nature of the data and hypotheses used differs from the one used here. The purpose of abduction is not to recover particular links, but rather to enumerate all possible additional links in prior networks to connect a target to a source node. The solutions are highly hypothetical, the inference process is not driven by experimental data and the logic-based inference about concurrent gene regulations works under the default assumption that an inhibitor always overpowers an activator.

The approach taken in the ARNI system follows a series of works based on physical network models (PNMs). PNMs aim to explain experimental observations on a template of protein-protein and TF-DNA interactions, by establishing causal chains between pairs of genes in such a way that the resulting information flow satisfies signal propagation principles. Many different formalisms have been explored for physical network reconstruction, including statistical scoring of active subnetworks [IDE 02], maximum-likelihood [YEA 04], linear programming [OUR 07], network flow optimization [YEG 09], and the Steiner tree approach [HUA 09, TUN 13]. The existing works on PNM approaches are limited in their ability to detect complex regulatory structures, see section 1.3.2.1, and the extent to which they can infer causal (signed-directed) networks. In [YEA 04, YEG 09, OUR 07], causal inference is restricted to source to target analysis which is not applicable to observational data where the source of perturbation is unknown. Methods that relax this assumption, [IDE 02, HUA 09, TUN 13], can only inferred unsigned undirected networks which have limited applicability for studying network dynamics and motifs. The formulation of the model, as presented in our ARNI approach, overcomes both of these limitations, allowing the inference of complex causal networks from observational and intervention data.

A unique contribution of our method is that, while improving topology inference, it constitutes a general logical framework that is elaboration tolerant, transparent to biologists and provides support for meta-level reasoning to test hypotheses. Recent

results have demonstrated how logic-based computational algorithms can be used to address problems such as modeling, analysis and revision of complex biological processes. Saez- Rodriguez *et al.* [SAE 09] developed CellNetOpt [TER 12], based on boolean and fuzzy logic, for optimizing signaling pathways against measurements of phosphorylation states. The problem of training logic-models of signaling pathways is revisited in [VID 12], which formulates the problem within the logic-based ASPenvironment and demonstrates a significant improvement on computation time. The work in [FAY 09] adapts an ASP framework for modeling cell cycle networks in yeast [DWO 08] to behave as a Boolean network. They conclude that the ASP framework outperforms Boolean networks both on expressiveness and scalability. In [GUZ 13], ASP is used to exhaustively characterize all possible boolean models of signaling pathways. The work in [GEB 10b] proposes a new library, called BioASP, to analyze biological networks with respect to a large amount of high-throughput biological data. BioASP expresses the sign consistency model presented in [SIE 06], which is closely related to our sign consistency constraints. Automated analysis tasks include detection and explanation of inconsistencies [GEB 08], computation of repairs and predictions [GEB 10a], and expansions of existing biological models [SCH 09]. Although very promising and effective on their computational tasks, none of these existing approaches can do de-novo topology inference. In all the above methods, biological networks are assumed given or known. More recently, abductive logic programming and ASP has been used in [LAZ 13, PAP 12], to analyze the effect experiments have on established networks and help biologists formulate new hypotheses and future experiments.

1.5.1. *Limitations and future work*

Incorporation of biological background knowledge is instrumental in overcoming the limitations of statistical approaches to gene network inference. This, however, introduces a bias in our ARNI approach towards interactions that have already been reported. Despite the growing body of available high throughput interaction assays providing a vast amount of interaction data, there are still unknown protein-protein or protein-DNA interactions that remain undiscovered. Our ARNI approach can be extended to complete networks with such previously unknown interactions. Logically and conceptually these can be accommodated very easily by introducing an additional abducible for interactive potentials and linking that to some constraints. Practically, we will need to drive the inference by integrating statistical approaches capable of detecting novel associations between genes, which might or might not correspond to physical interactions. Any statistical associations that cannot be explained in terms of paths in our approach, can be learned as new links in the networks. In this way we can perform the task of adding new links as proposed by [INO 13] but in a more realistic manner. In a recent work, [NOV 11] combines a Bayesian model describing modules of co-expressed genes and their corresponding transcription factors, and a physical interaction graph (undirected,unsigned), that links the transcription factors together.

Their approach is limited because of inability to include feedback loops, the PPI are unsigned-undirected, and the affected genes need to be preprocessed to clusters of genes.

ASP solvers identify all possible solutions, which can result in a large number of plausible networks. Our ARNI approach attempts to be as complete as possible in the initial step of computing the networks and then provides tools for the automatic generation of informative experiments that are most discriminatory over the inferred gene networks. Further work still needs to be done, in terms of definitions of informative experiments and revision operations, to formulate the revision task so to guarantee the entire process to eventually converge to a single gene network. The alternative networks can also be quantified probabilistically by incorporating our logic model in a framework that allows probabilistic abductive inference [TUR 13]. Such a framework would allow the representation of probabilistic abducibles, whose probability value can be interpreted as the the strength of the knowledge that led this link to be inferred. The higher the probability the higher the chance that the signed-directed link is true. Using these probabilities it could be possible to evaluate the probability of the inferred networks and therefore provide a means for performing model selection. Furthermore, using a BDD-based expectation maximization (EM) learning algorithm [INO 09] we could also learn the probabilities of the signed-directed interactions that would maximize the probability of each network (i.e. the success probability) and then use them to rank the networks in terms of their likelihood.

1.6. Conclusions

We have presented an approach, named ARNI, to logically model and automatically construct through abductive reasoning, regulatory gene networks from experimental data and background prior knowledge of gene interactions that might be known at the time. The main challenges in gene network inference are often considered to be the under-determined nature of the data (more parameters than data sets), the noisy and sparse nature of high throughput data and the complexity of network topologies. We have shown how ARNI makes key contributions to all three areas and through a series of evaluation experiments we have demonstrated the viability and potential of the approach.

The logical approach and nature of the constructed network models gives these models not only predictive power but also a high degree of versatility in their further development. They are easy to understand by the biologists and can be modularly changed either with new information that has become available or with hypotheses that the scientists want to examine before carrying out *in vitro* or *in vivo* experiments. We have shown how ARNI can be embedded within a general framework that supports automated scientific discovery where the validity of hypothesizing ideas can be examined and tested outside the laboratory.

This possibility of abstract analysis of potential ideas is central to the development of scientific theories and perhaps the main advantage of any logical approach to systems biology is that their high-level nature can facilitate this process of thought experimentation. Given the current descriptive and qualitative nature of much of biological knowledge, a logical formulation is well suited (compared to other formal approaches), for the development of tools that would allow the biologists to independently, i.e. without the continued help from computing experts, analyze their new scientific ideas and hypotheses before moving into the laboratory to test them. We envisage that it is possible to build an interface shell on top of ARNI that would provide such a tool for biologists who are studying regulatory cell networks.

1.7. Bibliography

[ALO 07] ALON U., "Network motifs: theory and experimental approaches", *Nature Reviews Genetics*, vol. 8, no. 6, pp. 450–461, 2007.

[BAR 04] BARABÁSI A.-L., OLTVAI Z.N., "Network biology: understanding the cell's functional organization", *Nature Reviews Genetics*, vol. 5, no. 2, pp. 101–113, 2004.

[BON 01] BONATTI P.A., "Reasoning with open logic programs", *Proceeding of International Conference on Logic Programming and Nonmonotonic Reasoning (LPNMR)*, pp. 147–159, 2001.

[BON 02] BONATTI P.A., "Abduction, ASP and open logic programs", *Proceeding of 9th International Workshop on Non-Monotonic Reasoning (NMR 02)*, pp. 184–190, 2002.

[DWO 08] DWORSCHAK S., GRELL S., NIKIFOROVA V.J., *et al.,* "Modeling biological networks by action languages via answer set programming", *Constraints*, vol. 13, nos. 1–2, pp. 21–65, 2008.

[FAI 07] FAITH J.J., HAYETE B., THADEN J.T., *et al.,* "Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles", *PLoS biology*, vol. 5, no. 1, p. e8, 2007.

[FAY 09] FAYRUZOV T., DE COCK M., CORNELIS C., *et al.,* "Modeling protein interaction networks with answer set programming", *Proceeding of IEEE International Conference on Bioinformatics and Biomedicine, BIBM'09*, IEEE, pp. 99–104, 2009.

[FRI 00] FRIEDMAN N., LINIAL M., NACHMAN I., *et al.,* "Using Bayesian networks to analyze expression data", *Journal of computational biology*, vol. 7, nos. 3–4, pp. 601–620, 2000.

[FRO 08] FRÖHLICH H., BEISSBARTH T., TRESCH A., *et al.,* "Analyzing gene perturbation screens with nested effects models in R and Bioconductor", *Bioinformatics*, vol. 24, no. 21, pp. 2549–2550, 2008.

[GEB 08] GEBSER M., SCHAUB T., THIELE S., *et al.,* "Detecting inconsistencies in large biological networks with answer set programming", *Logic Programming*, Springer, pp. 130–144, 2008.

[GEB 10a] GEBSER M., GUZIOLOWSKI C., IVANCHEV M., *et al.,* "Repair and prediction (under inconsistency) in large biological networks with answer set programming.", *proceeding of 12th International Conference on the Principles of Knowledge Representation and Reasoning (KR*), 2010.

[GEB 10b] GEBSER M., KONIG A., SCHAUB T., *et al.,* "The BioASP library: ASP solutions for systems biology", *22nd IEEE International Conference on Tools with Artificial Intelligence (ICTAI),* IEEE, vol. 1, pp. 383–389, 2010.

[GEB 12] GEBSER M., KAMINSKI R., KAUFMANN B., *et al., Answer Set Solving in Practice*, Synthesis Lectures on Artificial Intelligence and Machine Learning, Morgan and Claypool Publishers, 2012.

[GEL 88] GELFOND M., LIFSCHITZ V., "The stable model sematics for logic programming", *Proceedings of International Conference of Logic Programming*, pp. 1070–1080, 1988.

[GEN 04] GENTLEMAN R.C., CAREY V.J., BATES D.M., *et al.,* "Bioconductor: open software development for computational biology and bioinformatics", *Genome Biology*, vol. 5, no. R80, 2004.

[GUZ 13] GUZIOLOWSKI C., VIDELA S., EDUATI F., *et al.,* "Exhaustively characterizing feasible logic models of a signaling network using answer set programming", *Bioinformatics*, vol. 29, no. 18, pp. 2320–2326, 2013.

[HE 09] HE F., BALLING R., ZENG A.-P., "Reverse engineering and verification of gene networks: principles, assumptions, and limitations of present methods and future perspectives", *Journal of Biotechnology*, vol. 144, no. 3, pp. 190 – 203, 2009.

[HUA 09] HUANG S.S.C., FRAENKEL E., "Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks", *Science Signaling*, vol. 2, no. 81, pp. ra40–ra40, July 2009.

[IDE 02] IDEKER T., OZIER O., SCHWIKOWSKI B., *et al.,* "Discovering regulatory and signalling circuits in molecular interaction networks", *Bioinformatics*, vol. 18, no. suppl 1, pp. S233–S240, 2002.

[INO 09] INOUE K., SATO T., ISHIHATA M., *et al.,* "Evaluating abductive hypotheses using an EM algorithm on BDDs", *IJCAI*, pp. 810–815, 2009.

[INO 13] INOUE K., DONCESCU A., NABESHIMA H., "Completing causal networks by meta-level abduction", *Machine Learning*, vol. 91, pp. 239–277, 2013.

[IRR 10] IRRTHUM A., WEHENKEL L., GEURTS P. *et al*, "Inferring regulatory networks from expression data using tree-based methods", *PLoS One*, vol. 5, no. 9, p. e12776, 2010.

[JOS 10] JOSHI A., VAN PARYS T., PEER Y.V., *et al.,* "Characterizing regulatory path motifs in integrated networks using perturbational data", *Genome Biology*, vol. 11, no. 3, p. R32, 2010.

[KAK 90] KAKAS A.C., MANCARELLA P., "Database updates through abduction", *VLDB '90: Proceedings of the 16th International Conference on Very Large Data Bases*, Morgan Kaufmann Publishers Inc., San Francisco, CA. pp. 650–661, 1990.

[KAK 00] KAKAS A.C., MICHAEL A., MOURLAS C., "ACLP: abductive constraint logic programming", *Journal of Logic Programming*, vol. 44, nos. 1–3, pp. 129–177, 2000.

[KAK 01] KAKAS ANTONIS C., VAN NUFFELEN B., DENECKER M., "A-system : problem solving through abduction", *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, IJCAI, inc and AAAI, Morgan Kaufmann Publishers, Inc, vol. 1, pp. 591–596, 2001. Available at URL: http://www.cs.kuleuven.ac.be/cgi-bin-dtai/publ_info.pl?id=34862

[KIN 04] KING R.D., WHELAN K.E., JONES F.M., *et al.,* "Functional genomic hypothesis generation and experimentation by a robot scientist", *Nature*, vol. 427, no. 6971, pp. 247–252, 2004.

[KUF 12] KÜFFNER R., PETRI T., TAVAKKOLKHAH P., *et al.,* "Inferring gene regulatory networks by ANOVA", *Bioinformatics*, vol. 28, no. 10, pp. 1376–1382, 2012.

[LAN 08] LANGFELDER P., HORVATH S., "WGCNA: an R package for weighted correlation network analysis", *BMC bioinformatics*, vol. 9, no. 1, p. 559, 2008.

[LAZ 13] LAZAROU S., KAKAS A.C., NEOPHYTOU C., *et al.*, "Automated Scientific Assistant for Cancer and Chemoprevention", *Artificial Intelligence Applications and Innovations*, Springer Berlin Heidelberg, pp. 96–109, 2013.

[MAR 06] MARGOLIN A., NEMENMAN I., BASSO K., *et al.,* "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context", *BMC Bioinformatics*, vol. 7, no. 1, p. S7, 2006.

[MAR 09] MARBACH D., SCHAFFTER T., MATTIUSSI C., *et al.,* "Generating realistic in silico gene networks for performance assessment of reverse engineering methods", *Journal of Computational Biology*, vol. 16, no. 2, pp. 229–239, 2009.

[MAR 10] MARBACH D., PRILL R.J., SCHAFFTER T., *et al.,* "Revealing strengths and weaknesses of methods for gene network inference", *Proceedings of the National Academy of Sciences*, National Acad Sciences, vol. 107, no. 14, pp. 6286–6291, 2010.

[MAR 12a] MARBACH D., COSTELLO J.C., KUFFNER R., *et al.,* "Wisdom of crowds for robust gene network inference", *Nature Methods*, vol. 9, pp. 796–804, 2012.

[MAR 12b] MARBACH D., SCHAFFTER T., MATTIUSSI C., *et al.*, http://www.the-dream-project.org/category/challengesdream/dream5, 2012.

[MCI 94] MCILWRAITH S., "Generating tests using abduction", *Proceeding of 4th International Conference on Principles of Knowledge Representation and Reasoning, (KR 94)*, Morgan Kaufmann, pp. 449–460, 1994.

[MEN 06] MENDOZA L., "A network model for the control of the differentiation process in Th cells", *Biosystems*, vol. 84, no. 2, pp. 101–114, Elsevier, 2006.

[NOV 11] NOVERSHTERN N., REGEV A., FRIEDMAN N., "Physical module networks: an integrative approach for reconstructing transcription regulation", *Bioinformatics*, vol. 27, no. 13, pp. i177–i185, 2011.

[OUR 07] OURFALI O., SHLOMI T., IDEKER T., *et al.,* "SPINE: a framework for signaling-regulatory pathway inference from cause-effect experiments", *Bioinformatics*, vol. 23, no. 13, pp. i359–i366, July 2007.

[PAP 05] PAPATHEODOROU I., KAKAS A., SERGOT M., "Inference of gene relations from microarray data by abduction", *Logic Programming and Nonmonotonic Reasoning*, Springer Berlin Heidelberg, vol. 3662, pp. 389–393, 2005.

[PAP 12] PAPATHEODOROU I., ZIEHM M., WIESER D., *et al.,* "Using answer set programming to Integrate RNA expression with signalling pathway information to infer how mutations affect ageing", *PloS one*, vol. 7, no. 12, p. e50881, 2012.

[PRI 10] PRILL R., MARBACH D., SAEZ-RODRIGUEZ J., *et al.,* "Towards a rigorous assessment of systems biology models: the DREAM3 challenges", *PLoS ONE*, vol. 5, p. e9202, 2010.

[RAY 08] RAY O., BRYANT C.H., "Inferring the function of genes from synthetic lethal mutations", *Proceding of International Conference on Complex, Intelligent and Software Intensive Systems, CISIS,* IEEE, pp. 667–671, 2008.

[RAY 10] RAY O., WHELAN K., KING R., "Logic-based steady-state analysis and revision of metabolic networks with inhibition", *Proceeding of International Conference on Complex, Intelligent and Software Intensive Systems (CISIS 10),* IEEE, pp. 661–666, 2010.

[ROT 13] ROTIVAL M., PETRETTO E., "Leveraging gene co-expression networks to pinpoint the regulation of complex traits and disease, with a focus on cardiovascular traits", *Brief Funct Genomics*, vol. 13, pp. 66–78, 2013.

[SAE 09] SAEZ-RODRIGUEZ J., ALEXOPOULOS L.G., EPPERLEIN J., *et al.,* "Discrete logic modelling as a means to link protein signalling networks with functional analysis of mammalian signal transduction", *Molecular Systems Biology*, vol. 5, no. 1, p. 331, 2009.

[SCH 05] SCHÄFER J., STRIMMER K., "An empirical Bayes approach to inferring large-scale gene association networks", *Bioinformatics*, vol. 21, no. 6, pp. 754–764, 2005.

[SCH 09] SCHAUB T., THIELE S., "Metabolic network expansion with answer set programming", *Logic Programming*, Springer Berlin Heidelberg, vol. 5649, pp. 312–326, 2009.

[SCH 11] SCHAFFTER T., MARBACH D., FLOREANO D., "GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods", *Bioinformatics*, vol. 27, pp. 2263–2270, 2011.

[SIE 06] SIEGEL A., RADULESCU O., LE BORGNE M., *et al.,* "Qualitative analysis of the relation between DNA microarray data and behavioral models of regulation networks", *Biosystems*, vol. 84, no. 2, pp. 153–174, 2006.

[STO 07] STOLOVITZKY G., MONROE D., CALIFANO A., "Dialogue on reverse-engineering assessment and methods: the DREAM of high-throughput pathway inference", *Annals of the New York Academy of Sciences*, vol. 1115, pp. 11–22, 2007.

[TAM 06] TAMADDONI-NEZHAD A., CHALEIL R., KAKAS A., *et al.,* "Application of abductive ILP to learning metabolic network inhibition from temporal data", *Machine Learning*, Springer, vol. 64, nos. 1–3, pp. 209–230, 2006.

[TER 12] TERFVE C., COKELAER T., HENRIQUES D., *et al.,* "CellNOptR: a flexible toolkit to train protein signaling networks to data using multiple logic formalisms", *BMC Systems Biol*ogy, vol. 6, no. 1, p. 133, 2012.

[TRA 09] TRAN N., BARAL C., "Hypothesizing about signaling networks", *Journal of Applied Logic*, vol. 7, no. 3, pp. 253–274, 2009.

[TUN 13] TUNCBAG N., BRAUNSTEIN A., PAGNANI A., *et al.,* "Simultaneous reconstruction of multiple signaling pathways via the prize-collecting steiner forest problem", *Journal of Computational Biology*, vol. 20, no. 2, pp. 124–136, February 2013.

[TUR 13] TURLIUC C.-R., MAIMARI N., RUSSO A., *et al.,* "On minimality and integrity constraints in probabilistic abduction", *LPAR*, pp. 759–775, 2013.

[VID 12] VIDELA S., GUZIOLOWSKI C., EDUATI F., *et al.,* "Revisiting the training of logic models of protein signaling networks with ASP", *Computational Methods in Systems Biology*, pp. 342–361, 2012.

[WER 08] WERHLI A.V., HUSMEIER D., "Gene regulatory network reconstruction by Bayesian integration of prior knowledge and/or different experimental conditions", *Journal of bioinformatics and computational biology*, vol. 6, no. 03, pp. 543–572, 2008.

[YEA 04] YEANG C.-H., IDEKER T., JAAKKOLA T., "Physical network models", *Journal of Computational Biology*, vol. 11, nos. 2–3, pp. 243–262, 2004.

[YEG 04] YEGER-LOTEM E., SATTATH S., KASHTAN N., *et al.,* "Network motifs in integrated cellular networks of transcription–regulation and protein–protein interaction", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 16, pp. 5934–5939, 2004.

[YEG 09] YEGER-LOTEM E., RIVA L., SU L.J., *et al.,* "Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity", *Nature Genetics*, vol. 41, no. 3, pp. 316–323, February 2009.