

A Finite Model Theory for Biological Hypotheses

Stephen Racunas, Christopher Griffin and Nigam Shah
Penn State University
University Park, PA 16802
{sar147, cxg286, nhs109}@psu.edu

Abstract

We have designed and implemented a set of software tools for the composition and evaluation of hypotheses about gene regulation in biological systems. Our software uses a unified formal grammar for the representation of both diagram-based and text-based hypotheses. The objective of this paper is to show how to use this grammar as the basis for an effective logic for specifying hypotheses about biological systems in precise model-theoretic terms. To accomplish this, we take inspiration from inflationary extensions to fixed point logics and define a new type of logic: a deflationary logic for describing the effects of experiments upon models of biological systems. We present results that characterize decidability, satisfiability, and inflationary/deflationary properties of this logic. We formally define what it means for a set of assertions to be discoverable under this new logic, and show that our software generates discoverable queries. Thus, we lay the groundwork for a formal treatment of machine-aided experimental design under the conceptual framework we have developed for our hypothesis evaluation software.

1. Introduction

The representation and evaluation of hypotheses about biological processes is a challenging task because of the varying levels of detail at which biological processes can be considered and the variety of data types available [1]. To address this challenge, we have developed the Hypothesis-Space framework [7], a contradiction-based hypothesis testing framework designed to aid in the discovery of regulatory mechanisms.

Our Hypothesis-Space framework performs data fusion at the logical level [8]. Thus, we can incorporate heterogeneous information from many diverse sources in a principled manner [7]. (This had heretofore been a stumbling block for biological systems modeling efforts [6].) The Hypothesis-Space framework includes an event-based

grammar for representing biological processes and data from different experimental sources at varying levels of resolution [8]. We wish to use this grammar to support qualitative reasoning about biological systems.

Biologists speak of a “model” of a biological system [9]. By this, they typically mean a description of the biological entities and interactions present in the system, expressed in terms of diagrams and descriptive text [3]. In this paper, we show how to design a logic for specifying the concept of a biological model in precise model-theoretic terms.

Inspired by inflationary extensions to fixed point logics [4], we define and introduce a *deflationary* logic to be used for describing the effects of experiments upon models of biological systems. We then characterize this logic using techniques from finite model theory and present results on decidability, satisfiability, discoverability, and inflationary/deflationary operators.

2. Language and logic

2.1. Language

Definition 2.1. A *language* is a triple $\langle \mathcal{F}, \mathcal{R}, \mathcal{C} \rangle$ where:

1. \mathcal{F} is a set of function symbols f , each with positive integer arity n_f ,
2. \mathcal{R} is a set of relation symbols R , each with non-negative integer arity n_R , and
3. \mathcal{C} is a set of constant symbols c .

Remark 2.2. A 0-ary relation is a proposition.

Remark 2.3. Given a language L , we shall assume that we have available an infinite reserve of variables \mathcal{V} to use in function symbols and relation symbols.

Definition 2.4. The set of *L-terms* is the smallest set \mathcal{T} such that

1. $c \in \mathcal{T}$, where c is a constant of language L ,

2. $x \in \mathcal{T}$, where x is a variable in \mathcal{V} , and
3. $f(t_1, \dots, t_n) \in \mathcal{T}$, where f is an n -ary function and t_1, \dots, t_n are L -terms.

2.2. Logic

Definition 2.5. A *logic* is a formal set of rules for constructing formulas from a language.

Definition 2.6. Suppose that L is a language. An L -formula is defined inductively:

1. If R is an n -ary relation and t_1, \dots, t_n are L -terms, then $R(t_1, \dots, t_n)$ is an (atomic) L -formula.
2. If φ and ψ are formulas, then $\varphi * \psi$ is an L -formula, where $*$ $\in \{\wedge, \vee, \implies, \iff\}$.
3. If φ is an L -formula, then $\neg\varphi$ is an L -formula.

We will use the first-order predicate calculus with quantifier free formulas. We will allow n -ary predicates with constant symbols and propositional connectives. (This is equivalent to the propositional calculus, since each predicate with constant symbols is a proposition.)

The study of syntax is the study of logical proofs.

Definition 2.7. A *proof* is a finite sequence $\varphi_0, \varphi_1, \dots, \varphi_n$ such that φ_i is obtained from an axiom or is obtained from $\varphi_0, \dots, \varphi_{i-1}$ through the use of a deduction rule.

We have constructed an axiom set \mathbf{A}_B from facts that are generally accepted as true by the experimental community [5] for a specific biological system (galactose metabolism in *S. cerevisiae*). Please refer to www.hybrow.org for details.

The most common deduction rule is *modus ponens*, given as:

$$\frac{\varphi, \varphi \implies \psi}{\psi};$$

i.e., given formulas φ and $\varphi \implies \psi$, we can deduce ψ . Although this is the most common rule, other deduction rules are also used.

Definition 2.8. Suppose that φ is a formula. We say ψ can be deduced from φ (written $\varphi \vdash \psi$) if we can construct a proof ending with ψ that uses only φ and the axioms.

Definition 2.9. A formula φ in which no variables occur or in which all variables are bound is called a *sentence*.

Definition 2.10. Let S be a set of sentences. The *theory* of S , written $\text{Th}(S) = \{\varphi : S \vdash \varphi\}$, is the set of sentences that can be obtained from S using the deduction rules. Equivalently, the theory of S is the smallest set of formulas containing S and closed under \vdash .

Definition 2.11. We say that a set of sentences is *consistent* if there is no sentence φ such that $S \vdash \varphi$ and $S \vdash \neg\varphi$.

Semantics is the application of meanings to formulas and the study of formulas as they refer to structures.

Definition 2.12. Suppose that L is a language. A *model* (or L -structure) is a tuple $\mathcal{M} = \langle M, \mathcal{F}^{\mathcal{M}}, \mathcal{R}^{\mathcal{M}}, \mathcal{C}^{\mathcal{M}} \rangle$, where

1. M is a set called the universe,
2. $\mathcal{F}^{\mathcal{M}}$ is a set of functions $f^{\mathcal{M}} : M^{n_f} \rightarrow M$, where n_f is the arity of the function symbol $f \in L$,
3. $\mathcal{R}^{\mathcal{M}}$ is a set of relations $R^{\mathcal{M}} : R^{\mathcal{M}} \subseteq M^{n_R}$, where n_R is the arity of $R \in L$, and
4. $\mathcal{C}^{\mathcal{M}}$ is a set of constants in M .

We say that a model \mathcal{M} satisfies a certain sentence φ if φ is true in \mathcal{M} under assignment of constants to elements of the model. (The Tarskian definition of truth is applied in this instance, because it fully defines truth in terms of structures. A full description of Tarski's truth definition can be found in [2].) When a sentence φ is true in \mathcal{M} , we write $\mathcal{M} \models \varphi$ (read " \mathcal{M} satisfies φ ").

Definition 2.13. A set of sentences S is said to be *satisfiable* if there is a model \mathcal{M} such that $\mathcal{M} \models S$. (The set of models that satisfy a given set of sentences S is denoted $\text{Mod}(S)$.)

Remark 2.14. For the propositional calculus, a set of sentences is consistent if and only if it is satisfiable. This is called soundness and completeness of a logic.

For biological systems, we are almost certain that the universe will be subject to extensions and changes. Thus, satisfiability is not as important as in mathematical systems. Of greater interest is something we have chosen to call the "discoverability" property.

Definition 2.15. A set of sentences $\{\varphi_i : i \in X \subset \mathbb{N}\}$ is *discoverable* if there is a model \mathcal{M} such that $\exists \varphi \in \text{Th}(\mathbf{A}_B \cup \varphi_i)$ such that $\neg\varphi \in \text{Th}(\mathcal{M})$; *i.e.*, there is a model satisfying statements that contradict things that the sentences (together with the axioms) imply.

Fact 2.16. Our hypothesis grammar is capable of generating a logic over a system-specific language. (See www.hybrow.org for details.)

Remark 2.17. If our deduction system is to be biologically useful, we cannot use the simplification $\neg\neg\varphi \equiv \varphi$ since it is not necessarily true in experimental systems. If experimental data forces the falsity of a certain hypothesis $\neg\varphi$, it is *not* necessarily the case that φ is true.

2.3. Decidability

Theorem 2.18. *If \mathcal{L} is a logic and L is a finite language, then it is decidable whether a sequence of logical and extra-logical symbols φ is a formula of $\mathcal{L}[L]$.*

Proof. Apply a simple induction on the complexity of φ using the definitions of the formulas in the logic. \square

Corollary 2.19. *It is decidable whether or not a given hypothesis in the biological language we pose is well-formed; i.e., whether or not it is a structurally valid hypothesis.*

3. Model checking and hypotheses

A model checking problem has two inputs: a structure and a formula.

Definition 3.1. Given a language L , an L -structure \mathcal{M} and an L -formula φ , solving the *model checking problem* involves deciding whether φ is true in \mathcal{M} .

A closely related problem is that of query evaluation.

Definition 3.2. Given a structure \mathcal{M} and a formula $\varphi(x)$ with free variables x , solving the *query evaluation problem* involves computing the relation defined by φ on \mathcal{M} ; i.e., the set $\varphi^{\mathcal{M}} := \{a \in M : \mathcal{M} \models \varphi(a)\}$.

Remark 3.3. The evaluation problem for a formula with k free variable on a structure with n elements reduces to n^k model checking problems.

Biological model building is facilitated by experimentation. There is not, *a priori*, “one” model of \mathbf{A}_B , and each possible model $\mathcal{M} \in \text{Mod}(\text{Th}(\mathbf{A}_B))$ is a different explanation of the biological system that is consistent with the known facts. Hence, we wish to link discoverability to experimentation.

Definition 3.4. A sentence φ is *experimentally discoverable* if there exists an experiment which can show that φ is false.

Remark 3.5. Model construction may be difficult or impossible, depending upon the logic used to construct the formula φ .

Fact 3.6. Suppose that \mathcal{M} is a biological model. It is decidable whether φ is discoverable in \mathcal{M} if φ is propositional over the language of a biological system.

Remark 3.7. Though discoverability is decidable because of the decidability of satisfiability in the propositional calculus, it is still difficult to decide. (In fact, satisfiability is one of the archetypal NP complete problems.)

Each hypothesis H consists of a number of sentences composed according to a grammar G_H and using terms from the universe M . These sentences describe a theory $\text{Th}(H)$.

Let $V = \text{Th}(H) \cap \text{Th}(A_B)$ be the portion of the hypothesis that is valid *a priori*. Let R be a set of evaluation rules for the biological system. (Please refer to www.hybrow.org for examples of our evaluation rules and for details about how the evaluation rules are applied to real-world data.) Let $X = \text{Th}(R(D)) \cap \text{Th}(H)$ be the portion of the hypothesis contradicted by the rules as applied to data set D . In order for our framework to aid in automated experimental design, we desire everything *not* in the validated or contradicted sets to be discoverable.

Theorem 3.8. *Sentences not in $V \cup X$ are discoverable.*

Proof. G_H generates propositional sentences. \square

4. Deflationary logic

Ebbinhaus and Flum discuss inflationary fixed point logic (IFP) extensions to existing logics. In this section, we introduce the notion of *deflationary* logic and relate it to the role of experiments in the biological sciences.

Let \mathcal{M} be a finite model. A mapping $F : \mathcal{P}(M) \rightarrow \mathcal{P}(M)$ is said to be inflationary if $F(Z) \supseteq Z$ for all $Z \subseteq M$. If we begin with the empty set, we can form a sequence where $F_0 = \emptyset$ and $F_{n+1} = F(F_n)$. If there is some n such that $F_{n+1} = F_n$, then F is said to have a fixed point and we call this fixed point F_∞ .

Let φ be a formula with variables x_1, \dots, x_k (and possibly parameters) and let X be a relation symbol that occurs positively (i.e., is not negated) in φ . Let \mathcal{M} be a model. Then we define

$$F^\varphi(R) = \{a_1, \dots, a_k \in M : \mathcal{M} \models \varphi(a_1, \dots, a_k, R)\},$$

$R \subseteq M$. We can similarly define $(F^\varphi)' = F^{X(x_1, \dots, x_k) \vee \varphi}$, that is:

$$(F^\varphi)'(R) = \{a_1, \dots, a_k \in M : \mathcal{M} \models \varphi(a_1, \dots, a_k, R) \vee R(a_1, \dots, a_k)\}.$$

(Note: $R(a_1, \dots, a_k)$ simply means $a_1, \dots, a_k \in R$.)

From this, we can define a new logical operation, namely $[\text{Inf}_{X(\bar{x})\varphi}](\bar{t})$, where $\bar{t} = t_1, \dots, t_k$ are terms (possibly with parameters from a model). Then we have the following semantics:

$$\mathcal{M} \models [\text{Inf}_{X(x_1, \dots, x_k)\varphi}](t_1, \dots, t_k) \iff \bar{t}^{\mathcal{M}} \in (F^\varphi)'_\infty$$

In an experimental context, it is not appropriate to frame the investigative process in terms of “proofs” of “biological theorems.” The paradigm of the scientific method insists that the function of experimental work is to contradict and disprove hypotheses. This notion is at odds with the traditional role of mathematics, which is to prove new theorems given a set of assumptions.

We attempt to reconcile this issue by introducing the concept of deflationary logic. Let \mathcal{L} be a logic (first-order predicate, modal etc.) and let L be a language describing a biological system. Biological data can only disprove hypotheses (to a certain level of confidence). If \mathcal{M} is a (validated) L -structure in the logic \mathcal{L} , then $T = \text{Th}(\mathcal{M}) \supset \text{Th}(\mathbf{A}_B)$ is the set of all hypotheses true in the model \mathcal{M} . Essentially, \mathcal{M} is a “view of the world” and T is the set of all things true in that world. If φ is a sentence in the language L that is disproved by an experiment, then this gives rise to a new theory $T^* \subseteq T$, where

$$\psi \in T^* \iff \psi \in T \wedge \psi \not\vdash \varphi \wedge \psi \notin \text{Th}(\varphi).$$

Let $F : \mathcal{P}(T) \rightarrow \mathcal{P}(T)$. We shall say that F is deflationary if for all $S \subseteq T$, $F(S) \subseteq S$. A fixed point of F is a set such that $F(S) = S$. We denote such a fixed point by T_∞ . We shall say that F is trivial if \emptyset is the unique fixed point of F . We can create a sequence:

$$T_0, T_1, \dots, T_\infty,$$

where $T_0 = \text{Th}(\mathbf{A}_B)$, $T_1 = F(T_0)$ etc. If we restrict our attention to only finite models of theories, then we can relate the deflationary operator F to deflationary and inflationary operators on the models.

Suppose that \mathcal{M} is a finite model of a theory T . Then $\text{Th}(\mathcal{M})$ is finite by necessity, and the application of F to $\text{Th}(\mathcal{M})$ will lead to a new theory $T^* \subseteq \text{Th}(\mathcal{M})$.

Theorem 4.1. *If F is a deflationary operator and \mathcal{M} is a finite model of a theory T , then there is a finite model \mathcal{M}^* such that $\mathcal{M}^* \models T^*$ and $|\mathcal{M}^*| = |\mathcal{M}|$ and $\mathcal{M} = \mathcal{M}^*$.*

Proof. Without loss of generality, suppose that T^* contains one less sentence, φ , than T . If φ is atomic, then it has the form $R(a_1, \dots, a_k)$ and hence we can construct a model \mathcal{M}^* from \mathcal{M} by removing elements a_1, \dots, a_k from $R^{\mathcal{M}}$. Suppose that $\varphi = \neg R(a_1, \dots, a_k)$. Then we simply add a_1, \dots, a_k to $R^{\mathcal{M}}$ to form \mathcal{M}^* . Now suppose that φ is conjunctive, i.e., $\varphi = \psi \wedge \eta$. Then by induction, we can create a model \mathcal{M}_1^* for the removal of formula φ and a model \mathcal{M}_2^* for the removal of η . Since we assume that ψ and η are consistent, we can form the model \mathcal{M}^* from either \mathcal{M}_1^* or \mathcal{M}_2^* by treating \mathcal{M}_1^* (or \mathcal{M}_2^* , respectively) as the base model and removing the appropriate sentence. If $\varphi = \psi \vee \eta$, then as before we have a model \mathcal{M}_1^* when we remove ψ and a model \mathcal{M}_2^* when we remove η . Let \mathcal{M}^*

be either of these, or the model obtained in the conjunctive case. Now, since each formula can be written in disjunctive normal form ($\bigvee_i \bigwedge_j R_{i,j}$, where $R_{i,j}$ is either a atomic formula or its negation), we can form \mathcal{M}^* for any formula. \square

Remark 4.2. We see from the proof that if φ is a positive formula atomic formula, then the relations of \mathcal{M} are smaller than the relations of \mathcal{M}^* . Conversely, if φ is not positive, then the relations of \mathcal{M}^* are larger than those of \mathcal{M} . Conjunctions of positive and negative formula will inflate or deflate relations accordingly.

Using this approach, we can define inflationary and deflationary operators $F_{\mathcal{M}}^\uparrow$ and $F_{\mathcal{M}}^\downarrow$ on $\mathcal{P}(\mathcal{M})$. Let T and \mathcal{M} be as above. Let us assume that F removes exactly $H = \{\psi : \psi \vdash \varphi\}$ for some special φ . Since we are considering only quantifier free predicate formulae, we can assume that H is composed of atomic formula and their negations, since removing a formula of the form $\alpha \wedge \beta$ is identical to removing both α and β . H is countable and can be separated into the disjoint union H^+ , the positive formulae and H^- the negative formulae. Let $F_{\mathcal{M}}^\uparrow$ map subsets of \mathcal{M} to subsets of \mathcal{M} corresponding to the removal of formula in H^- . Likewise, let $F_{\mathcal{M}}^\downarrow$ be defined for H^+ . In this way, alterations to \mathcal{M} correspond to mixed inflationary and deflationary actions.

5. Expressiveness

Definition 5.1. Let \mathcal{L}_1 and \mathcal{L}_2 be logics. Then, $\mathcal{L}_1 \leq \mathcal{L}_2$ (read “ \mathcal{L}_1 is at most as expressive as \mathcal{L}_2 ”) if for every τ and every sentence $\varphi \in \mathcal{L}_1[\tau]$, there is a sentence $\psi \in \mathcal{L}_2[\tau]$ such that $\text{Mod}(\varphi) = \text{Mod}(\psi)$.

Lemma 5.2. *If F_1 and F_2 are two deflationary operators on T , then $F_1 \circ F_2(T) = F_2 \circ F_1(T)$.*

Proof. Without loss of generality, suppose F_1 removes ψ_1 and F_2 removes ψ_2 . Then,

$$\begin{aligned} F_1(F_2(T)) &= F_2(F_1(T)) \\ &= \{\psi : \psi \in T \wedge \psi \not\vdash \psi_1 \wedge \psi \not\vdash \psi_2 \wedge \psi \notin \text{Th}(\psi_1, \psi_2)\} \end{aligned}$$

\square

If we assume the consistency of the initial theory T and are given \mathcal{M} , a model of T , then $F^\infty(T)$ yields a theory and a corresponding unique model \mathcal{M}^* (built from \mathcal{M}), such that $\mathcal{M}^* \models F^\infty(T)$ and the relations of T are obtained by the inflation and deflation of relations in \mathcal{M} . This logic is like an inflationary logic, but incorporates both inflation and deflation of the relations.

Unfortunately, the expressive power of this logic depends entirely on the deflationary operator F and the resulting theory $F^\infty(T)$. Hence it cannot be determined

without a complete knowledge of \mathcal{M}^* . Let IDFP be the *inflationary-deflationary fixed point logic* generated from the *quantifier free* logic QF we have defined.

Thus we deduce:

Theorem 5.3. *If every relation of \mathcal{M}^* is larger than or equal to the relations of \mathcal{M} , then $\text{QF}(\text{IDFP}) \geq \text{QF}$. Conversely, if every relation of \mathcal{M}^* is smaller than or equal to the relations of \mathcal{M} , then $\text{QF}(\text{IDFP}) \leq \text{QF}$. Otherwise, the two logics are incomparable.*

We now also show that we can safely extend \mathcal{L} by adding constants as we determine new biological elements:

Theorem 5.4. *Suppose that L is a language and $L^* = L \cup \{c_1, \dots, c_n\}$. Suppose \mathcal{M} is an L -structure and φ is an L -sentence with unknown place holders u_1, \dots, u_n . If $\mathcal{M} \models \varphi$, then \mathcal{M} is also an L^* -structure and $\mathcal{M} \models \varphi^*$ when we replace u_1, \dots, u_n with c_1, \dots, c_n .*

Proof. Simply let the interpretation of c_1, \dots, c_n be the original interpretation of u_1, \dots, u_n . \square

6. Future work

A logic equipped with a hierarchy of temporal and causal operators would be of greater biological utility. We will determine the most concise logic of this type that is appropriate for working with models of gene regulatory systems.

If a mapping is established between the syntax of a biological logic and a set of experimental procedures, it may be possible to derive an experiment to decide whether φ is discoverable in a given model (or class of models). If the time complexity required for the derivation and completion of the experiment is less than that required to decide discoverability in a given model, then there exists a “biological algorithm” that solves this NP-complete problem quickly. We will seek to determine conditions under which this is true.

But before we investigate these directions, we will build upon the foundations laid down in this work to formalize a rich structure for the space of all possible hypotheses that can be composed for a particular biological system. We will explore the possibility of using Stone Space concepts to introduce a topology on the space of hypotheses. Given an ontology for a biological system, an associated logic, experimental data, and a set of rules for interpreting the data, we will define a neighborhood structure and perturbation operators for moving through and for modifying the Hypothesis Space. In short, we wish to develop operations and transformations on models and on the Hypothesis Space that reflect the ways in which experimental biologists alter their working models as they progress toward greater understanding of gene regulatory systems. These extensions to the Hypothesis Space framework will allow us to construct further software tools to aid in the experimental design process.

7. Conclusions

We have presented a new type of logic designed to facilitate computer aided experimental design under the Hypothesis-Space framework. We have demonstrated decidability, satisfiability, and inflationary/deflationary properties for this logic. We have defined what it means for a set of assertions to be “discoverable” under this logic, and have demonstrated that the hypothesis composition grammar used by our “Hypothesis Browser” software prototype generates queries that are discoverable.

These results help construct a foundation for the formal treatment of machine-aided experimental design under the conceptual framework we have developed for our hypothesis evaluation software. In future work, we will build upon this foundation and develop further structure to the space of expressible hypotheses for a given biological system. These extensions to the Hypothesis Space framework will facilitate the development of further software tools that streamline the process of experimental design.

References

- [1] R. B. Altman and S. Raychaudhuri. Whole-genome expression analysis: challenges beyond clustering. *Curr Opin Struct Biol*, 11(3):340–7, 2001.
- [2] J. Bell and M. Machover. *A Course in Mathematical Logic*. North-Holland, Amsterdam, 1977.
- [3] D. L. Cook, J. F. Farley, and S. J. Tapscott. A basis for a visual language for describing, archiving and analyzing functional models of complex biological systems. *Genome Biol*, 2(4), 2001.
- [4] H. Ebbinghaus and J. Flum. *Finite Model Theory*. Springer-Verlag, New York, 1999.
- [5] D. Lohr, P. Venkov, and J. Zlatanova. Transcriptional regulation in the yeast gal gene family: a complex genetic network. *Faseb J*, 9(9):777–87, 1995.
- [6] M. Peleg, I. Yeh, and R. B. Altman. Modelling biological processes using workflow and petri net models. *Bioinformatics*, 18(6):825–37, 2002.
- [7] S. Racunas, N. Shah, I. Albert, and N. Fedoroff. Hybrow: A prototype system for computer-aided hypothesis evaluation. *Bioinformatics*, 20(suppl. 1):i1–i8, 2004.
- [8] S. Racunas, N. Shah, and N. Fedoroff. A conceptual framework for hypothesis testing and evaluation. *Proc. IEEE CSB*, 2:634–638, 2003.
- [9] O. Wolkenhauer. Systems biology: The reincarnation of systems theory applied in biology? *Briefings in Bioinformatics*, 2(3):258–270, 2001.