PHENOTYPES AND PHILOSOPHY: FORMALISING SYSTEMS BIOLOGY HYPOTHESES FOR A ROBOT SCIENTIST

Alexander H. Gower[ab], Filip Kronström[ab], Daniel Brunnsåker[ab], Ievgeniia A. Tiukova[ac] and Ross D. King[abde]

[a] Chalmers University of Technology, Gothenburg, Sweden
[b] University of Gothenburg, Gothenburg, Sweden
[c] KTH Royal University of Technology, Stockholm, Sweden
[d] Cambridge University, Cambridge, United Kingdom
[e] Alan Turing Institute, London, United Kingdom

Abstract

An important part of the scientific process is the formation of hypotheses. A clear statement of a hypothesis in unambiguous language is the basis of good experimental design. Our challenge is to develop a method to record hypotheses in formal languages, with the precision required for robot scientists to execute experiments, and also that each hypothesis makes a clear statement about *Saccharomyces cerevisiae* biology.

Existing ontologies are used to describe scientific concepts in hypotheses. For example, we use terms from the Ascomycete Phenotype Ontology (APO) to describe observable traits of *S. cerevisiae*. In its current form, comparative statements are implicitly in reference to changes in genotype, so we extend APO to allow for arbitrary comparison in hypothesis statements.

Changes in phenotype, genotype, and logical implications are described using concept inclusions in description logic. This ontology is being integrated into existing ontologies that we have developed to specify experimental design and execution with a robot scientist[1].

Phenotype

Observable traits of an organism. Defined in the context of the yeast *Saccharomyces cerevisiae* as:

> "Features of ascomycete fungal cells, cultures, or colonies that can be detected, observed, measured, or monitored."
>
> – *Ascomycete Phenotype Ontology (APO)*

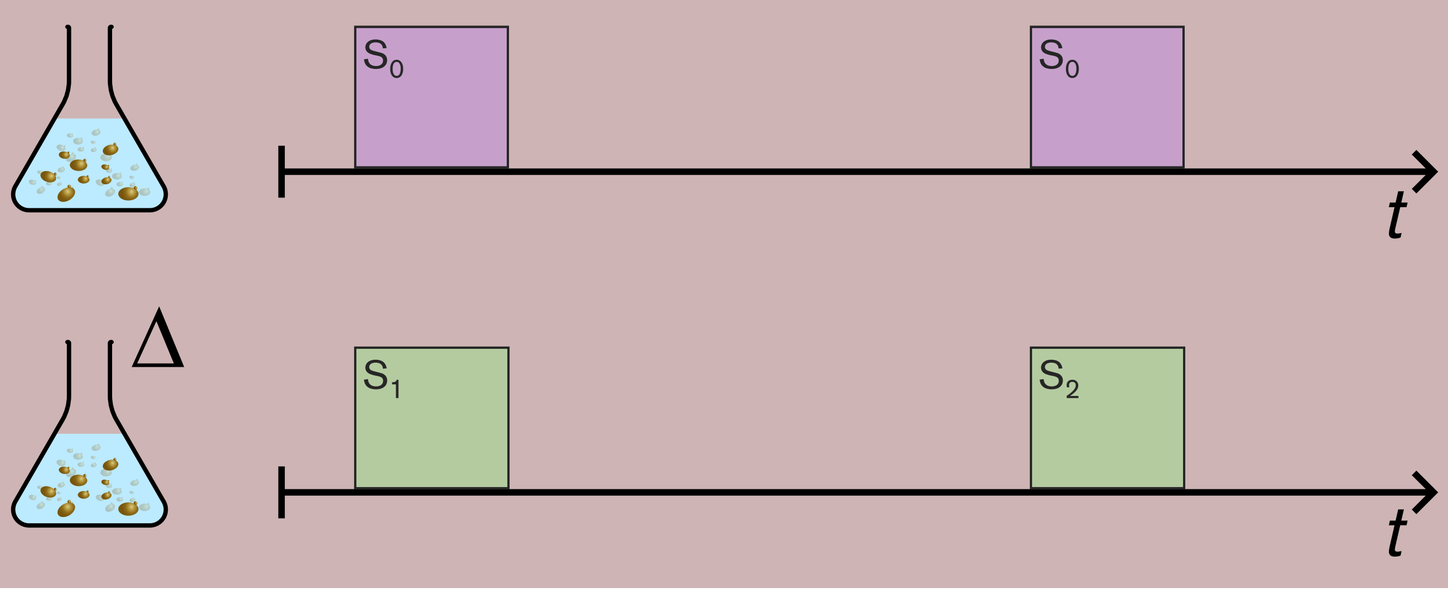
Implication in hypotheses

Scientific hypotheses seek to provide a theoretical explanation for a phenomenon. A common form of hypothesis is a **relational statement** couched in **relative change**.

"cells with **lower** intracellular alanine levels **exhibit decreased** heat resistance."

The relational statement is most useful when it is a causal implication. And to examine a causal implication we need to consider the temporal aspect of the phenomenon.

For *S. cerevisiae,* we predict and examine how the organism state changes over time.



Experiments

Statistical or logical tests based on empirical data collected from experiments are used to evaluate a hypothesis.

Robot scientists can execute different forms of experiments, but common to all is that the yeast will be cultivated over time.

Control experiments will provide the reference data for the test.

Implication over time can be expressed in different ways, for example with S_1 and S_2 two states:

$$\forall e\ \forall t\ \exists t'\ (t < t' \wedge (T(S1(e), t) \rightarrow T(S2(e), t')))$$

Hypothesis	"Cells with lower intracellular methionine levels exhibit increased acid pH resistance."

Hypothesis expressed using concept inclusions

// Defining base state

$S_0 \sqsubseteq$ organismState
$S_0 \sqsubseteq \exists$stateHasObservable.O_{01}
$S_0 \sqsubseteq \exists$stateHasObservable.O_{02}

$O_{01} \sqsubseteq$ 'chemical compound accumulation'
$O_{01} \sqsubseteq \exists$accumulationOfChemical.methionine

$O_{02} \sqsubseteq$ 'acid pH resistance'

// Defining perturbation state

$S_1 \sqsubseteq$ organismState
$S_1 \sqsubseteq \exists$stateHasObservable.O_{11}

$O_{11} \sqsubseteq$ 'chemical compound accumulation'
$O_{11} \sqsubseteq \exists$accumulationOfChemical.methionine
$O_{11} \sqsubseteq \exists$decreasedComparedTo.$O_{01}$

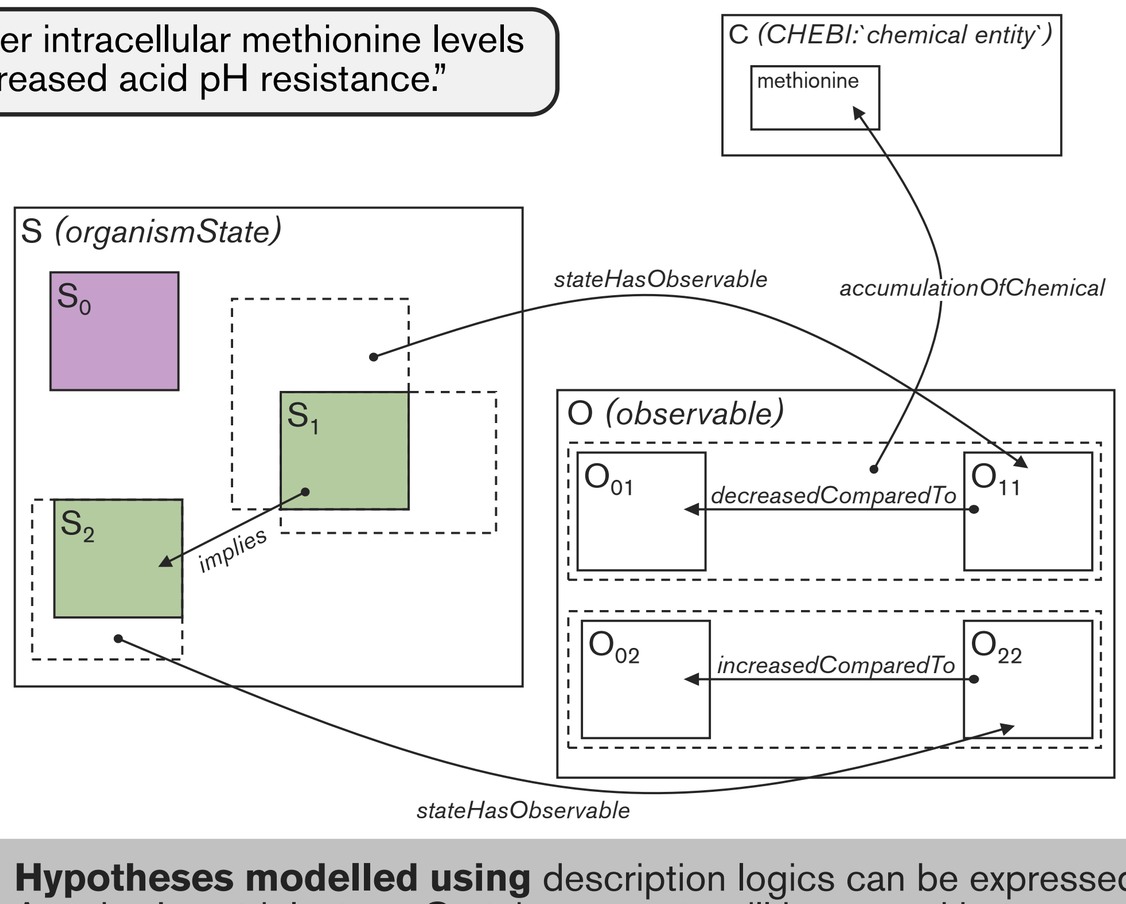
// Defining result state

$S_2 \sqsubseteq$ organismState
$S_2 \sqsubseteq \exists$stateHasObservable.O_{22}

$O_{22} \sqsubseteq$ 'acid pH resistance'
$O_{22} \sqsubseteq \exists$increasedComparedTo.$O_{02}$

//Hypothesis

$S_1 \sqsubseteq \exists$implies.S_2



Reference states

Differences in **phenotype** are measured against a reference state, or control.

Commonly, this is a **wild type** strain (a standard **genotype**, or genetic configuration) cultivated in a standard condition.

APO defines phenotypes using the **observable** class. Differences are defined implicitly with reference to the **wild type**. So implicit in APO definitions is that the induced variation from the reference state is only in the genotype.

We use description logics to describe concepts that allow us to specify arbitrary reference states allowing for much richer forms of hypothesis.

Hypotheses modelled using description logics can be expressed neatly in RDF triples. We store these in an Apache Jena triplestore. Certain concepts will be reused between many hypotheses. Storing and retrieving hypotheses like this prevents duplication, allowing increased efficiency and potential for semantic connections.

Future/outlook

We are testing this modelling framework with a robot scientist that has automated experimental protocols, and AI software generating new hypotheses. More can be done to couple the formalised hypotheses directly to experimental ontologies[1]. Linking the hypotheses directly to the tests and empirical data, as well as modelling uncertainty in conclusions, is an exciting future direction for this work.

[1] Reder, G. K. et al. Genesis-DB: a database for autonomous laboratory systems. Bioinformatics Advances 3, vbad102 (2023).
[2] Gower, A. H. et al. The Use of AI-Robotic Systems for Scientific Discovery. Preprint at https://doi.org/10.48550/arXiv.2406.17835 (2024).

This work is partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.