The Robot Scientist Genesis: Abduction for Metabolic Modelling
Alexander H. Gower <gower@chalmers.se>, Daniel Brunnsäker <danbru@chalmers.se> Filip Kronstrom <filipko@chalmers.se>, Gabriel Reder <reder@chalmers.se>, levgenia Tiukova <tiukova@chalmers.se>, Ronald S. Reiserer <ron.reiserer@Vanderbilt.Edu>, John Wikswo <john.wikswo@vanderbilt.edu>, Konstantin Korovin <Konstantin.Korovin@manchester.ac.uk>, Ross D. King <rossk@chalmers.se>

Background
At Chalmers in Sweden we are building a next generation Robot Scientist "Genesis" (King et al., 2004, 2009; Williams et al., 2015). Our goal is to demonstrate that the Robot Scientist Genesis can investigate an important area of science a thousand times more efficiently (in terms of cost and money) than human scientists. This is an extreme challenge for AI as the number of experiments to plan and coordinate is several orders of magnitude more than any previous work. Achieving this goal will involve advances in automated hypothesis formation (how best to utilise background biological knowledge and models in ML, etc.), automated experiment generation (how best to optimise gain of information with cost and time constraints), laboratory robotic control, and scientific data analysis. The application domain of Genesis is to develop a systems biology model of yeast (Saccharomyces cerevisiae), that is both more detailed and more accurate at predicting experimental results than any in existence (Coutant et al., 2019). Modelling yeast, the 'model' eukaryote, is central to the future of medicine, agriculture, and biotechnology. The foundation of Genesis is a micro-fluidic system with 1000 computer-controlled μ-bioreactors (co-developed in Vanderbilt University, USA). Achieving this will be a step-change in laboratory automation as most biological labs have <10 chemostats. These μ-bioreactors are being ingratiated with ion-flow mass-spectroscopy (to measure metabolites at speed) and RNA-seq (to measure RNA expression levels).

Scientific Discovery and Abduction
Metabolic network models represent the cellular biochemistry of an organism and the related action of enzymatic genes; such models which seek to integrate knowledge from the entire organism are known as genome-scale metabolic models (GEMs). GEMs require significant research effort to develop, which is expensive with regards to time, money and physical resource. The scientific discovery problem for improvement of GEMs is to add or remove knowledge (reactions, protein rules, et.) such that model quality is increased. Model quality in GEMs is multi-faceted—desirable properties of a model include predictive power (how well deductions using the model match experimental data), network coverage (the extent to which different parts of metabolism are represented in the model) and parsimony, and there is evidence to suggest that there are trade-offs between different desirable properties (Heavner & Price, 2015). Examples of improvements to GEMs of S. cerevisiae including improving annotation, removing noise from low-confidence components, and adding reactions to eliminate so-called “dead-end” compounds.

Automated techniques are one promising way to make scientific discoveries within systems biology at the scale and pace required for automated science. We have constructed a logical theory of yeast...
metabolic pathways using curated GEMs as the expert knowledge source. First-order logic enables the rich expression of knowledge about biological processes. Mechanisms such as reactions, enzyme catalysis and gene regulation can be expressed independently of specific genes, species or enzymes. Model improvement consists broadly of three stages: (1) hypothesise refinements to the model; (2) convert hypotheses and resultant model to a format suitable for simulation; and (3) perform an evaluation informed by experimental evidence and internal consistency (Thiele & Palsson, 2010). We use the automated theorem proving software iProver (Korovin, 2008) for each of these stages.

We show that by conducting deductive inference on the GEM-based logical theories using iProver we can predict the growth/no-growth phenotype in *S. cerevisiae* for combinations of genotype and environment (growth medium). In cases where these deductions are empirically incorrect we are using iProver to abduce hypotheses consisting of combinations of compounds whose presence would result in an error correction for several genes. These abductions represent gaps in the model, possibly from a missing reaction that produces the compound. For each of the 56 NGG errors (predicted non-growth, observed growth) in the single-gene deletion task, we abduced hypotheses using iProver. In total we generated 2,649 unique hypotheses; some hypotheses would result in an error correction for several genes. We filtered these hypotheses based on biological knowledge to produce 765 more probable hypotheses to evaluate. For these we apply two criteria for assessing the merit of each hypothesis. Firstly, we use the reactions activated in the proof found by iProver for each hypothesis to constrain simulations. Secondly, we repeat the single-gene essentiality prediction task using the initial theory with the hypothesis added. We are now integrating this hypotheses testing with empirical experiments in Genesis.

**References**


