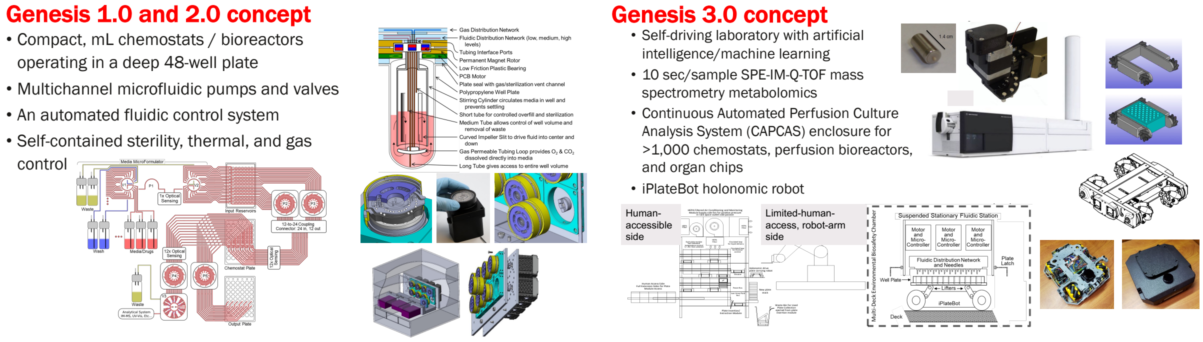
The Robot Scientist Genesis: Abduction for Metabolic Modelling

Alexander H. Gower <gower@chalmers.se>, Daniel Brunnsåker <danbru@chalmers.se> Filip Kronstrom <filipkro@chalmers.se>, Gabriel Reder <reder@chalmers.se>, Ievgeniia 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).



Genesis 1.0 and 2.0 concept

- Compact, mL chemostats / bioreactors operating in a deep 48-well plate
- Multichannel microfluidic pumps and valves
- An automated fluidic control system
- Self-contained sterility, thermal, and gas control

Genesis 3.0 concept

- Self-driving laboratory with artificial intelligence/machine learning
- 10 sec/sample SPE-IM-Q-TOF mass spectrometry metabolomics
- Continuous Automated Perfusion Culture Analysis System (CAPCAS) enclosure for >1,000 chemostats, perfusion bioreactors, and organ chips
- iPlateBot holonomic robot

Gas Distribution Network
Fluidic Distribution Network (low, medium, high levels)
Tubing Interface Ports
Permanent Magnet Rotor
Low Friction Plastic Bearing
PCB Motor
Plate seal with gas/sterilization vent channel
Polypropylene Well Plate
Stirring Cylinder circulates media in well and prevents settling
Short tube for controlled overfill and sterilization
Medium Tube allows control of well volume and removal of waste
Curved Impeller Slit to drive fluid into center and down
Gas Permeable Tubing Loop provides O2 & CO2 dissolved directly into media
Long Tube gives access to entire well volume
Media MicroFormulator
Waste
V1
P1
1x Optical Sensing
V2
Input Reservoirs
P2
P3
12-to-24 Coupling Connector: 24 in, 12 out
Wash
Media/Drugs
12x Optical Sensing
Chemostat Plate
12x Optical Sensing
P4
P5
V3
Waste
Analytical System IM-MS, UV-Vis, Etc...
Output Plate
Human-accessible side
Limited-human-access, robot-arm side
Multi-Deck Environmental Biosafety Chamber
Suspended Stationary Fluidic Station
Motor and Micro-Controller
Motor and Micro-Controller
Motor and Micro-Controller
Fluidic Distribution Network and Needles
Plate Latch
Well Plate
Lifters
iPlateBot
Deck

Figure 1. (a) A test module for Genesis-Lab. (b) Overall structure of Genesis-Lab.

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, etc.) 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

Coutant, A., Roper, K., Trejo-Banos, D., Bouthinon, D., Carpenter, M., Grzebyta, J., Santini, G., Soldano, H., Elati, M., Ramon, J., Rouveirol, C., Soldatova, L. N., & King, R. D. (2019). Closed-loop cycles of experiment design, execution, and learning accelerate systems biology model development in yeast. *Proceedings of the National Academy of Sciences*, *116*(36), 18142–18147. https://doi.org/10.1073/pnas.1900548116

Heavner, B. D., & Price, N. D. (2015). Comparative Analysis of Yeast Metabolic Network Models Highlights Progress, Opportunities for Metabolic Reconstruction. *PLOS Computational Biology*, *11*(11), e1004530. https://doi.org/10.1371/journal.pcbi.1004530

King, R. D., Rowland, J., Oliver, S. G., Young, M., Aubrey, W., Byrne, E., Liakata, M., Markham, M., Pir, P., Soldatova, L. N., Sparkes, A., Whelan, K. E., & Clare, A. (2009). The automation of science. *Science*, *324*(5923). https://doi.org/10.1126/science.1165620

King, R. D., Whelan, K. E., Jones, F. M., Reiser, P. G. K., Bryant, C. H., Muggleton, S. H., Kell, D. B., & Oliver, S. G. (2004). Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, *427*(6971), Article 6971. https://doi.org/10.1038/nature02236

Korovin, K. (2008). IProver – An Instantiation-Based Theorem Prover for First-Order Logic (System Description). In A. Armando, P. Baumgartner, & G. Dowek (Eds.), *Automated Reasoning* (Vol. 5195, pp. 292–298). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-71070-7_24

Thiele, I., & Palsson, B. Ø. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocols*, *5*(1), Article 1. https://doi.org/10.1038/nprot.2009.203

Williams, K., Bilsland, E., Sparkes, A., Aubrey, W., Young, M., Soldatova, L. N., De Grave, K., Ramon, J., de Clare, M., Sirawaraporn, W., Oliver, S. G., & King, R. D. (2015). Cheaper faster drug development validated by the repositioning of drugs against neglected tropical diseases. *Journal of the Royal Society, Interface*, *12*(104), 20141289. https://doi.org/10.1098/rsif.2014.1289