The Robot Scientist Genesis: Abduction for Metabolic Modelling

AAAI Spring Symposium
Computational Approaches to Scientific Discovery
San Francisco, 28 March 2023

Alexander H. Gower, Konstantin Korovin, Daniel Brunnsåker, Filip Kronstrom, Gabriel K. Reder, Ievgeniia A. Tiukova, Ronald S. Reiserer, John Wikswo, Ross D. King
The original discovery problem

How we formulated the problem in computational terms
What data and knowledge we provided to our system
How we represented the system's inputs and outputs
The space of candidate models that the system searched
What criteria it used to evaluate candidate models
How we interpreted results that the system generated
Metabolic modelling *Saccharomyces cerevisiae*

- Yeast is the model eukaryote
- Exist tools to conduct experiments (e.g. CRISPR/Cas9)
- Cell factory
Metabolic modelling *Saccharomyces cerevisiae*

“The ultimate goal of genome-scale metabolic network reconstruction in the future is to have a well-annotated network including all parts of the metabolism without any missing reactions or gaps; however it is not yet possible due to incomplete knowledge of the yeast metabolism.”

— Österlund et. al (2012)
Metabolic modelling *Saccharomyces cerevisiae*

**Genome scale metabolic model**
- Some quantities can be measured directly
- Others are abstractions (e.g. metabolic fluxes)
- Common approach is to encode biological knowledge as constraints
  - Either from observed experiments or
  - Biophysical knowledge

iFF708
Reactions: 1145
Metabolites: 825
Genes: 708

iMH805/77
R: 1149
M: 646
G: 750

iN800
R: 1446
M: 1013
G: 800

Yeast 4.0
R: 2030
M: 1481
G: 924

iMM904b
R: 1575
M: 1228
G: 904

iTO977
R: 1038
M: 636
G: 672

Yeast 1.0
R: 1761
M: 1168
G: 888

iAZ900
R: 1597
M: 1398
G: 900

iIN800
R: 1446
M: 1013
G: 800

Yeast 8
R: 3949
M: 2680
G: 1133

GECKO technique

Yeast 1.0 + lipid
+ suitable for simulation

3 cellular compartments*

8
Transcription regulation integrated

3
Lipid metabolism (143 reactions)

8
Yeast 1.0 + lipid

8


3 cellular compartments*

8

3

8

8

8

8

3

3

8

8

3

8

8

First consensus model

Refinement of Yeast 5

Separation of stoichiometry constraints from met. model

Corrections to Yeast 6

Improvements to lipid metabolism

* - cytosol, mitochondria, extracellular space
** - above + nucleus, endoplasmic reticulum, golgi, vacuole and peroxisome
Model improvement

- Model reduction
- Model expansion (new annotation)
  - Regulatory interaction
  - Assign gene to known reaction
  - Predict missing reaction
- Condition-specific effects
- Compartmentalisation (split model over parts of cell)
- New mathematical rules (enzymatic rate equation)
  \[ v_i = k_{cat,i} \cdot \sum E_i \cdot \rho_i \]
- New constraint mechanism (e.g. introducing enzymes explicitly into equations, GECKO)

Orth & Palsson (2010)

Sanchez, Zhang et. al. (2017)
Humans and machines working together

- Model reduction
- Model expansion (new annotation)
  - Regulatory interaction
  - Assign gene to known reaction
  - Predict missing reaction
- Condition-specific effects
- Compartmentalisation (split model over parts of cell)
- New mathematical rules (enzymatic rate equation)
  \[ v_i = k_{cat,i} \cdot \sum E_i \cdot \rho_i \]
- New constraint mechanism (e.g. introducing enzymes explicitly into equations, GECKO)

Algorithms exist or are being developed for these methods

Currently proposed by human scientists – require a level of abstraction
How to compare models

Many possible metrics one could use:

- genomic coverage;
- overlap of annotated metabolites;
- predictive ability for single gene essentiality;
- biomass production prediction;
- ...

Good models have:

- explanatory power
- predictive power
- consistency across contexts
- consistency with other scientific models

- It is difficult to find a single metric that can summarise a model's quality
- Among *S. cerevisiae* models there is evidence of tradeoffs between predictive accuracy and gene network coverage (Heavner and Price, 2015)
- Models are often developed for specific applications
The original discovery problem

How we formulated the problem in computational terms

What data and knowledge we provided to our system

How we represented the system's inputs and outputs

The space of candidate models that the system searched

What criteria it used to evaluate candidate models

How we interpreted results that the system generated
Logical inference

<table>
<thead>
<tr>
<th>Logical Inference</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>induction</td>
<td>allows us to generalise models from data</td>
</tr>
<tr>
<td>deduction</td>
<td>given a theory what conclusions can we draw</td>
</tr>
<tr>
<td>abduction</td>
<td>how can we “fix” the theory to be consistent with empirical data?</td>
</tr>
</tbody>
</table>

Active learning

Machine learning paradigm where the learning agent has agency over the selection of the next data to learn from — analogous to the scientific method
Genesis flow

Systems Biology Theory → Generate Hypotheses

Simulate Experiments → Metabolomics Analysis

Genome Runs Experiments in Laboratory → Transcriptomics Analysis

Predicted Phenotype → Refine the Theory

Observed Phenotype → Improved Systems Biology Theory
Background: Genesis

• Scalable automated biological experimentation
• Small volume chemostat cultivations – vision is for thousands of parallel experiments
• AI-driven laboratory machine
• Measurements via high-throughput metabolomics and transcriptomics

John Wikswo group VIIBRE
The original discovery problem

How we formulated the problem in computational terms

What data and knowledge we provided to our system

How we represented the system's inputs and outputs

The space of candidate models that the system searched

What criteria it used to evaluate candidate models

How we interpreted results that the system generated
Constructing a logical theory of metabolism

- Background knowledge is encoded in the curated genome-scale metabolic models (GEMs)
- Each reaction in the GEM is translated to a set first-order logic clauses
- Clauses are written in conjugate normal form to produce theory
- Inference is performed using iProver: a theorem prover for first-order logic with support for arithmetical reasoning (Korovin, 2008)
- iProver has the efficiency required for a logical theory on this scale

### GEM (SBML)

<table>
<thead>
<tr>
<th>Reaction</th>
<th>6-phospho-D-gluconate + NADP(+) (\text{EC} 1.1.1.35) (\rightarrow) carbon dioxide + D-ribulose 5-phosphate + NADPH</th>
</tr>
</thead>
</table>

### Reaction identifier

- **r_0889**
- **phosphogluconate dehydrogenase**
- **stoichiometry**: \(s_{0340}[c] + s_{1207}[c] \rightarrow s_{0456}[c] + s_{0577}[c] + s_{1212}[c]\)
- **GPR**: YGR256W or YHR183W

### Logical Formuлаe

#### Reaction activation

\[
\text{rxn}(R) \leftarrow \text{met}(M_1, C_1) \land \cdots \land \text{met}(M_N, C_N) \\
\land \text{enz}(E^3)
\]

#### Reaction products

\[
\text{met}(M, C) \leftarrow \text{rxn}(R)
\]

#### Enzyme availability

\[
\text{enz}(E) \leftarrow \text{pro}(P_1^E) \lor \cdots \lor \text{pro}(P_M^E)
\]

#### Protein formation

\[
\text{pro}(P) \leftarrow \text{gn}(G_1^P) \lor \cdots \lor \text{gn}(G_N^P)
\]

#### Gene activation

\[
\text{gn}(G) 
\]

#### Metabolite availability

\[
\text{met}(M, C) 
\]

### Logical Theory Clauses

**Predicate** | **Natural language interpretation**
--- | ---
met\(\{2\} | “Metabolite X is present in cellular compartment Y.”
gn\{1\} | “Gene X is expressed.”
pro\{1\} | “Protein complex X is available (in every cellular compartment).”
enz\{1\} | “Isoenzyme category X is available.”
rxn\{1\} | “There is positive flux through reaction X.”
The original discovery problem
How we formulated the problem in computational terms
What data and knowledge we provided to our system
How we represented the system's inputs and outputs
The space of candidate models that the system searched
What criteria it used to evaluate candidate models
How we interpreted results that the system generated
Proof graph from iProver (simplified)

Glycolysis pathway (from https://pathway.yeastgenome.org/)
The original discovery problem
How we formulated the problem in computational terms
What data and knowledge we provided to our system
How we represented the system's inputs and outputs
The space of candidate models that the system searched
What criteria it used to evaluate candidate models
How we interpreted results that the system generated
**Approach: abduction opportunities**

- Some parts of the model are well-known due to the chemistry (reaction stoichiometry, protein formation)

- We seek to:
  a) learn rules about which enzymes catalyse which reactions
  b) identify possible missing reactions by finding compound presence that will repair a broken pathway

- Future work will be to learn rules for gene expression and activation

---

<table>
<thead>
<tr>
<th>Predicate</th>
<th>Natural language interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>met\1</td>
<td>“Metabolite X is present in cellular compartment Y.”</td>
</tr>
<tr>
<td>gn\1</td>
<td>“Gene X is expressed.”</td>
</tr>
<tr>
<td>pro\1</td>
<td>“Protein complex X is available (in every cellular compartment).”</td>
</tr>
<tr>
<td>enz\1</td>
<td>“Isoenzyme category X is available.”</td>
</tr>
<tr>
<td>rxn\1</td>
<td>“There is positive flux through reaction X.”</td>
</tr>
</tbody>
</table>

---

**Logical Theory Clauses**

1. $\text{rxn(r\_0889)} \leftarrow \text{met(s\_0340, c\_c)} \land \text{met(s\_1207, c\_c)} \land \text{enz(e\_r\_0889)}$
2. $\text{enz(e\_r\_0889)} \leftarrow \text{pro(p\_r\_0889\_1)} \lor \text{pro(p\_r\_0889\_2)}$
3. $\text{pro(p\_r\_0889\_1)} \leftarrow \text{gn(YGR256W)}$
4. $\text{pro(p\_r\_0889\_2)} \leftarrow \text{gn(YHR183W)}$
5. $\text{met(s\_0456, c\_c)} \leftarrow \text{rxn(r\_0889)}$
6. $\text{met(s\_0577, c\_c)} \leftarrow \text{rxn(r\_0889)}$
7. $\text{met(s\_1212, c\_c)} \leftarrow \text{rxn(r\_0889)}$

---

**Logical Formulae**

- Reaction activation: $\text{rxn(R)} \leftarrow \text{met(M\_1, C\_1) \land \cdots \land met(M\_N, C\_N)} \land \text{enz(E\_R)}$
- Reaction products: $\text{met(M, C)} \leftarrow \text{rxn(R)}$
- Enzyme availability: $\text{enz(E)} \leftarrow \text{pro(P\_1\_E) \lor \cdots \lor pro(P\_M\_E)}$
- Protein formation: $\text{pro(P)} \leftarrow \text{gn(G\_1\_P) \land \cdots \land gn(G\_P\_P)}$
- Gene activation: $\text{gn(G)}$
- Metabolite availability: $\text{met(M, C)}$

---

**Reaction**

\[
\begin{align*}
6 \text{-phospho-D-gluconate} & \; + \; \text{NADP}(+) & \text{EC} & 1.1.1.331 \\
\text{carbon dioxide} & \; + \; \text{D-ribulose 5-phosphate} & \; + \; \text{NADPH} \\
\rightarrow & \\
\end{align*}
\]
The original discovery problem
How we formulated the problem in computational terms
What data and knowledge we provided to our system
How we represented the system's inputs and outputs
The space of candidate models that the system searched
What criteria it used to evaluate candidate models
How we interpreted results that the system generated
Evaluating models: predicting single-gene essentiality

• Which of the ~6000 genes are necessary for healthy growth?
• We assume the presence of extracellular metabolites that correspond to YNB (yeast nitrogen base) plus glucose
• Systematically remove a gene from the model by negating the relevant clause $g(g_{to\_knock\_out})$ becomes $\sim g(g_{to\_knock\_out})$
• Then query for the essential metabolites

<table>
<thead>
<tr>
<th>Base GEM</th>
<th>Yeast 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of predictions (number of genes in model)</td>
<td>1068 (1150)</td>
</tr>
<tr>
<td>NG Recall (NGNG / NG)</td>
<td>0.16 (25/161)</td>
</tr>
<tr>
<td>NG Precision (NGNG / (NGNG + NGG))</td>
<td>0.31 (25/81)</td>
</tr>
<tr>
<td>GNG Rate (GNG / NG)</td>
<td>0.845 (136/161)</td>
</tr>
<tr>
<td>NGG Rate (NGG / G)</td>
<td>0.062 (56/907)</td>
</tr>
<tr>
<td>F1 score</td>
<td>0.207</td>
</tr>
</tbody>
</table>
Evaluating models: constraining flux balance simulations

Stoichiometric matrix - \( S \)

Fluxes - \( \nu \)

maximize \( \nu \in \mathbb{R}^n \)
subject to \( S\nu = 0, \)
\( \nu_i^{LB} \leq \nu_i \leq \nu_i^{UB}, \quad i = 1, \ldots, n. \)

"Is the pathway for production of pyruvate from a minimal glucose medium feasible?"

1.45e-13
distance from reference flux distribution

"Which compounds need to be added to a minimal glucose medium to enable production of malonyl CoA?"

6.51
distance from reference flux distribution

iProver

reference flux distribution

hypotheses

\( \text{omf(abundance_extras_001.raxiom.me(m_\text{glycinebione}_n_e))}. \)
The original discovery problem
How we formulated the problem in computational terms
What data and knowledge we provided to our system
How we represented the system's inputs and outputs
The space of candidate models that the system searched
What criteria it used to evaluate candidate models
How we interpreted results that the system generated
Interpreting results

• Can use automated theorem prover (iProver) for deduction but also abduction—has the efficiency required for models of this size
• Finding models is easy, finding good models is hard
• Using multiple deduction techniques can help check model consistency—bridge the divide
• Conflicting evidence in literature—e.g. pathway for L-arginine production
• Limits to what one can do with others’ data
Future work

• Hypothesis testing with Genesis platform
• Integration of metabolomics and transcriptomics measurements
• Learn rules for gene expression and regulation
• Integrate with signalling pathways
Acknowledgements

Konstantin Korovin
King Lab
Erik Bjurström
Daniel Brunnsåker
Filip Kronström
Praphapan Lasin
Gabriel Reder
Ievgeniia Tiukova
Ross D. King

Vanderbilt University
John Wikswo
Ron Reiserer
Clayton Britt
Greg Gerken
Kyle Hawkins
Dmitry Markov
Philip Samson
David Schaffer
Eric Spivey
Erik Werner

Larisa N. Soldatova
Thoughtworks
Rushikesh Halle
Vinay Mahamuni
Amit Patel
Harshal Hayatnagarkar

...and more!
## Areas of improvement

<table>
<thead>
<tr>
<th>Challenge - from Chen Y, Li F, Nielsen J (2022)*</th>
<th>Hypotheses</th>
<th>Techniques to test hypotheses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotation of the model</td>
<td>Correct formulae, charge etc. for metabolites; enzyme numbers; reaction directionality</td>
<td>Mass balance analysis; ??</td>
</tr>
<tr>
<td>Noise from low-confidence components</td>
<td>Removal of certain elements from model; additional dimension of model “explains” noise</td>
<td>; condition-specific evaluation techniques</td>
</tr>
<tr>
<td>“Dead-end” metabolites</td>
<td>Add reactions</td>
<td>Predictive accuracy of metabolic activity with/without reactions; thermodynamic balance analysis</td>
</tr>
<tr>
<td>Un- or poorly-annotated reactions (in particular transport reactions)</td>
<td>New or changed gene-protein rule</td>
<td>Comparing transcriptomics or proteomics data with metabolic activity; mutant strain cultivation</td>
</tr>
<tr>
<td>Changes to biomass equation itself</td>
<td>Coefficient change; variable addition or removal; condition-specific biomass equations</td>
<td>Comparative prediction analysis</td>
</tr>
<tr>
<td>Enzyme turnover rate estimation</td>
<td>Values for enzyme turnover rates</td>
<td>Experimentally measure enzyme levels; comparative prediction analysis</td>
</tr>
<tr>
<td>Integration of subcellular constraints</td>
<td>Reaction constraints e.g. within mitochondrion</td>
<td>Quantification of sub-cellular proteomes; ??</td>
</tr>
<tr>
<td>Integration of regulation mechanisms</td>
<td>More precise mathematical formulations of regulatory mechanisms; better models of currently understood mechanisms of regulation</td>
<td>Comparative prediction analysis; multi-omics analysis; mutant strain cultivation</td>
</tr>
</tbody>
</table>

Role of automation

“Note that all the computer-predicted additions should be recorded and require expert level verification.” — Yu et. Al (2022)

Automated laboratory processes and AI techniques bring:

1. Efficiency
2. Broader reasoning (as opposed to deep reasoning right now)
3. Precision and repeatability

I would summarise by saying the strength of automation (in scientific discovery for S. cerevisiae models) will be in embracing complexity by avoiding excessive simplification during reasoning and exploiting big datasets to extract small signals from large noise, particularly when it comes to testing condition-specific models.