Automating Chemistry and Biology using Robot Scientists

Ross D. King,
University of Manchester, ross.king@manchester.ac.uk
Winston Churchill as Philosopher
“When my metaphysical friends tell me that the data on which the astronomers made their calculations ... were necessarily obtained originally through the evidence of their senses, I say 'No'. They might, in theory at any rate, be obtained by automatic calculating machine set in motion by the light falling upon them without admixture of the human senses at any stage”

Winston Churchill (1930)
Three Philosophical Worlds

Matter

Consciousness

Ideas
Robot Scientist

Matter

Ideas
Winston Churchill and Scientific Advice to Government
Scientific Advice to Government

Frederick Lindemann

Patrick Blackett
Scientific Advice to Government

- For many years AI was not taken seriously by the scientific or political establishment.

- Now it is increasingly clear that AI will be one of the defining technologies of the 21\textsuperscript{st} century.

- It is likely that a number of you will give advice to Governments about AI.

- It is important that you get this advice correct.
Robot Scientists
The Concept of a Robot Scientist

Computer systems capable of originating their own experiments, physically executing them, interpreting the results, and then repeating the cycle.
Motivation: Philosophical

- What is Science?

- The question whether it is possible to automate scientific discovery seems to me central to understanding science.

- There is a strong philosophical position which holds that we do not fully understand a phenomenon unless we can make a machine which reproduces it.
Richard Feynman’s Blackboard

“What I cannot create, I do not understand”
Motivation: Technological

- Robot Scientists have the potential to increase the productivity of science. They can work cheaper, faster, more accurately, and longer than humans. They can also be easily multiplied.
  - Enabling the high-throughput testing of hypotheses.

- Robot Scientists have the potential to improve the quality of science.
  - by enabling the description of experiments in greater detail and semantic clarity.
The Complexity of Biological Systems

- Even simple “model” biological systems like that of *E. coli* and yeast are incredibly complicated.
- Thousands of genes, proteins, small-molecules, interacting together in complicated spatial temporal ways.
- Ockham's razor doesn't work - system evolved.

- Not enough PhDs in the world to disentangle these systems.
- Need help - Robot Scientists.
Motivation AI

■ Science is a wonderful test bed for AI.

■ Compare/Contrast with Chess
  – Small abstract world: 64 squares, 36 pieces.
  – Computers now play chess much better than the best humans: ELO 2,800 v ELO 3,350.
  – Computers can now make strikingly beautiful moves.

  – No special “magic” for intelligence: increased quantity of search made a qualitative difference.
Motivation AI

- Scientific problems are abstract, but involve the real-world.
- Scientific problems are restricted in scope – no need to know about “Cabbages and Kings”.
- Nature is honest – no malicious agents.
- Nature is a worthy object of our study.
- The generation of scientific knowledge is a public good.
Scientific Discovery
Meta-Dendral

Analysis of mass-spectrometry data.

Robot Scientist Timeline

1999-2004 Initial Robot Scientist Project
- Limited Hardware
- Collaboration with Douglas Kell (Aber Biology), Steve Oliver (Manchester), Stephen Muggleton (Imperial)

2004-2011 Adam Project – Yeast Functional Genomics
- Sophisticated Laboratory Automation
- Collaboration with Steve Oliver (Cambridge).
  King et al. (2009) Science, 324, 85-89

2008-2015 Eve Project – Drug Design for Tropical Diseases
Adam
The Application Domain

- Functional genomics
- In yeast (S. cerevisiae) ~15% of the 6,000 genes still have no known function.
- EUROFAN 2 made all viable single deletant strains.
- **Task** to determine the “function” of a gene by growth experiments.
Formalising the Problem

- Use logic programming to represent background knowledge: metabolism modelled as a directed labeled hyper-graph.
- Use abduction to infer new hypotheses:
  - Abductive logic programming.
  - Techniques from Bioinformatics.
- Use active learning to decide efficient experiments: cost of compounds and time.
- Use machine learning to decide meaning of experimental results.
The Experimental Cycle

- Background Knowledge
- Hypothesis Formation
- Analysis
- Final Theory
- Experiment(s) selection
- Robot
- Results Interpretation
Model v Real-World

Experimental Predictions

Logical Model

Biological System

Experimental Results
We have developed a logical formalism for modelling metabolic pathways (encoded in Prolog). This is essentially a directed labeled hyper-graph: with metabolites as nodes and enzymes as arcs.

If a path can be found from cell inputs (metabolites in the growth medium) to all the cell outputs (essential compounds) then the cell can grow.
Genome Scale Model of Yeast Metabolism

- It covers most of what is known about yeast metabolism.
- Includes 1,166 ORFs (940 known, 226 inferred).
- Growth if path from growth medium to defined end-points.
- State-of-the-art accuracy in predicting cell viability.

- Now integrated with Yeast 4.0.
The Experimental Cycle

- **Background Knowledge**
- **Hypothesis Formation**
- **Analysis**
- **Final Theory**
- **Experiment(s) selection**
- **Robot**
- **Results Interpretation**
Inferring Hypotheses

■ Science is based on the hypothetico-deductive method.

■ In the philosophy of science. It has often been argued that only humans can make the “leaps of imagination” necessary to form hypotheses.

■ In biology most hypothesis generation is abductive, not inductive.

■ Adam used abductive inference to infer missing arcs/labels in its metabolic graph - hypotheses. With these missing nodes Adam could then deductively infer (explain) the observed experimental results.
Automated Model Completion

Bioinformatics Database + FASTA32 PSI-BLAST → Model of Metabolism

Hypothesis Formation → Experiment Formation

REACTION

Gene Identification

Deduction
orthologous(Gene1, Gene2) → similar_sequence(Gene1, Gene2).

Abduction
similar_sequence(Gene1, Gene2) → orthologous(Gene1, Gene2).
The Experimental Cycle

Background Knowledge → Hypothesis Formation ← Analysis

Final Theory ← Experiment(s) selection → Robot → Results Interpretation
The Form of the Experiments

- Hypothesis 1: Gene YER152C codes for the enzyme the reaction: chorismate $\rightarrow$ prephenate.
- Hypothesis 2: Gene YGL060W codes for the enzyme the reaction: chorismate $\rightarrow$ prephenate.

These can be tested by:
- Growing YER152C$\Delta$ in environment +/- prephenate.
- Growing YGL060W$\Delta$ in environment +/- prephenate.
Phenylalanine, Tyrosine, and Tryptophan Pathways for *S. cerevisiae*
Inferring Experiments

Given a set of hypotheses we wish to infer an experiment that will efficiently discriminate between them

Assume:

■ Every experiment has an associated cost.
■ Each hypothesis has a probability of being correct.

The task:

■ To choose a series of experiments which minimise the expected cost of eliminating all but one hypothesis.
Active Learning

- In the 1972 Fedorov (Theory of optimal experiments) showed that this problem is in general intractable (NP complete).

- However, it can be shown that the problem is the same as finding an optimal decision tree; and it is known that this problem can be solved “nearly” optimally in polynomial time.

- We have shown that this strategy can outperform (get answer faster and cheaper) than simply choosing the cheapest experiment. Also better than humans on test problem.
The Experimental Cycle

Background Knowledge $\rightarrow$ Hypothesis Formation $\leftarrow$ Analysis

Consistent Hypotheses

Final Theory $\rightarrow$ Experiment(s) selection $\rightarrow$ Robot $\rightarrow$ Results Interpretation
Diagram of Adam
Adam in Action
Growth plates
Example Growth Curves

<table>
<thead>
<tr>
<th>Plate ID: jun09-titr7-expt001-plate001 Barcode 16248</th>
<th>Arginine as N source</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td><img src="image1.png" alt="Graphs" /></td>
</tr>
<tr>
<td>B</td>
<td><img src="image2.png" alt="Graphs" /></td>
</tr>
<tr>
<td>C</td>
<td><img src="image3.png" alt="Graphs" /></td>
</tr>
<tr>
<td>D</td>
<td><img src="image4.png" alt="Graphs" /></td>
</tr>
<tr>
<td>E</td>
<td><img src="image5.png" alt="Graphs" /></td>
</tr>
<tr>
<td>F</td>
<td><img src="image6.png" alt="Graphs" /></td>
</tr>
<tr>
<td>G</td>
<td><img src="image7.png" alt="Graphs" /></td>
</tr>
<tr>
<td>H</td>
<td><img src="image8.png" alt="Graphs" /></td>
</tr>
</tbody>
</table>
Growth curves
The Experimental Cycle

- Background Knowledge
- Hypothesis Formation
- Experiment(s) selection
- Robot
- Final Theory
- Analysis
- Results Interpretation
Qualitative to Quantitative

- The functions of most genes in *S. cerevisiae* that when deleted result in auxotrophy (no growth) have already been discovered.

- Most genes of unknown function only affect growth quantitatively.

- They may have slower growth (bradytrophs), faster growth, higher/lower biomass yield, etc.
The Experimental Cycle

- Background Knowledge
- Hypothesis Formation
- Analysis
- Experiment(s) selection
- Robot
- Final Theory
- Results Interpretation
Closing the Loop

- We have physically implemented all aspects of Adam.

- To the best of our knowledge Adam was the first AI system that can both explicitly form hypotheses and experiments, and physically do the experiments.
Discovery of Novel Science
Adam generated and confirmed twelve novel functional-genomics hypotheses concerning the identity of genes encoding enzymes catalysing orphan reactions in the metabolic network of the yeast *S. cerevisiae*.

Adam's conclusions have been manually verified using bioinformatic and biochemical evidence.
## Novel Scientific Knowledge

<table>
<thead>
<tr>
<th>Orphan Enzyme</th>
<th>Hypothesised Gene</th>
<th>Prob.</th>
<th>Acc.</th>
<th>No.</th>
<th>Existing Annotation</th>
<th>Dry</th>
<th>Wet</th>
</tr>
</thead>
<tbody>
<tr>
<td>glucosamine-6-phosphate deaminase (3.599.6)</td>
<td>YHR163W (SOL3)</td>
<td>&lt;10⁻⁴</td>
<td>97</td>
<td>8</td>
<td>'6-phosphogluconolactonase' ida</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>glutaminase (3.5.1.2)</td>
<td>YL033C (BCY1)</td>
<td>&lt;10⁻⁴</td>
<td>92</td>
<td>11</td>
<td>'cAMP-dependent protein kinase inhibitor' ida</td>
<td>✗</td>
<td>?</td>
</tr>
<tr>
<td>L-threonine 3-dehydrogenase (1.1.1.103)</td>
<td>YDL168W (SFA1)</td>
<td>&lt;10⁻⁴</td>
<td>83</td>
<td>6</td>
<td>'alcohol dehydrogenase' ida</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>purine-nucleoside phosphorylase (2.4.2.1)</td>
<td>YLR209C (PNP1)</td>
<td>&lt;10⁻⁴</td>
<td>82</td>
<td>11</td>
<td>'purine-nucleoside phosphorylase' ida</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>2-aminoadipate transaminase (2.6.1.39)</td>
<td>YGL202W (AR08)</td>
<td>&lt;10⁻⁴</td>
<td>80</td>
<td>3</td>
<td>'aromatic-amino-acid transaminase' ida</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>5,10-methenyltetrahydrofolate synthetase (6.3.3.2)</td>
<td>YER183C (FAU1)</td>
<td>&lt;10⁻⁴</td>
<td>80</td>
<td>4</td>
<td>'5,10-formyltetrahydrofolate cyclo-ligase' ida</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>glucose-6-phosphate deaminase (3.5.90.6)</td>
<td>YNR034W (SOL1)</td>
<td>&lt;10⁻⁴</td>
<td>79</td>
<td>2</td>
<td>'possible role in tRNA export'</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>pyridoxal kinase (2.7.1.35)</td>
<td>YPR121W (THI22)</td>
<td>&lt;10⁻⁴</td>
<td>78</td>
<td>1</td>
<td>'phosphomethylpyrimidine kinase' ida</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>mannitol-1-phosphate 5-dehydrogenase (1.1.1.17)</td>
<td>YPR073C</td>
<td>&lt;10⁻⁴</td>
<td>78</td>
<td>6</td>
<td>'putative mannitol dehydrogenase' ida</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1-acylglycerol-3-phosphate O-acetyltransferase (2.3.1.51)</td>
<td>YDL052C (SIC1)</td>
<td>0.0001</td>
<td>80</td>
<td>6</td>
<td>'1-acylglycerol-3-phosphate O-acetyltransferase' ida</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>glucose-6-phosphate deaminase (3.5.99.6)</td>
<td>YGR248W (SOL4)</td>
<td>0.0002</td>
<td>78</td>
<td>2</td>
<td>'6-phosphogluconolactonase' ida</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>maleylacetate isomerase (5.2.1.2)</td>
<td>YLL060C (GT2)</td>
<td>0.0003</td>
<td>76</td>
<td>3</td>
<td>'glutathione S-transferase' ida</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>sene O-acetyltransferase (2.3.1.30)</td>
<td>YJL218W</td>
<td>0.0005</td>
<td>78</td>
<td>2</td>
<td>'unknown function'</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>L-threonine 3-dehydrogenase (1.1.1.103)</td>
<td>YLR070C (XYL2)</td>
<td>0.0052</td>
<td>75</td>
<td>6</td>
<td>'xylitol dehydrogenase' ida</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2-aminoadipate transaminase (2.6.1.39)</td>
<td>YJL060W (BNA3)</td>
<td>0.0084</td>
<td>73</td>
<td>3</td>
<td>'kynurenine aminotransferase' ida</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>pyridoxal kinase (2.7.1.35)</td>
<td>YNR027W</td>
<td>0.0259</td>
<td>76</td>
<td>2</td>
<td>'involved in bud-site selection' ida</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>polyamine oxidase (1.5.3.11)</td>
<td>YMR020W (FMS1)</td>
<td>0.0289</td>
<td>78</td>
<td>4</td>
<td>'polyamine oxidase' ida</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>2-aminoadipate transaminase (2.6.1.39)</td>
<td>YER152C</td>
<td>0.0332</td>
<td>74</td>
<td>3</td>
<td>'uncharacterized'</td>
<td>-</td>
<td>✓</td>
</tr>
<tr>
<td>L-aspartate oxidase (1.4.3.16)</td>
<td>YJL045W</td>
<td>0.1300</td>
<td>72</td>
<td>1</td>
<td>'succinate dehydrogenase isozyme' ida</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>purine-nucleoside phosphorylase (2.4.2.1)</td>
<td>YLR017W (MEU1)</td>
<td>0.1421</td>
<td>72</td>
<td>6</td>
<td>'methylthioadenosine phosphorylase' ida</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
Eve
Parasitic Diseases targeted

Malaria

Shistosomaisis

Leishmaniasis

Chagas
Why Tropical Diseases?

- Millions of people die of these diseases, and hundreds of millions of people suffer infection.

- It is clear how to cure these diseases – kill the parasites.

- They are “neglected”, so avoid competition from the Pharmaceutical industry.
Formalising the Problem

- Use graphs and standard chemoinformatic methods to represent background knowledge - the use of relations is planned.

- Uses induction (quantitative structure activity relationship – QSAR learning) to infer new hypotheses.

- Use active learning to decide efficient experiments, and econometric model to decide what compounds to test.
Automating Early Drug Development

Synthetic Biology → Robot Scientist

Assay Design → Library screen → Hit Confirmation → Learn and Test QSAR → Lead Compound
Eve’s Automation of Pipeline

- Standard library screening is brute force:
  - Eve uses intelligent screening

- In the standard “pipeline” the 3 processes are not integrated.
- In Eve automated and integrated.
Eve’s Hardware

Highlights of Eve's hardware:

- Acoustic liquid handling
- High throughput 384 well plates
- Two industrial robot arms
- Automated 60x microscope
- Liquid handlers, fluorescence readers, barcode scanners, dry store, incubator, tube decapper...
Intelligent v Brute-force Screening

- We wished to compare our AI based screening against the standard brute-force approach: “begin at the beginning and go on till you come to the end: then stop” (Lewis Carroll).

- While simple to automate standard screening is slow and wasteful of resources, since every compound in the library is tested. It is also unintelligent, as it makes no use of what is learnt during screening.

- Use money to decide.
The Economics of Intelligent Screening

\[ \Delta \text{Utility of Eve} = \sum_{1}^{Nm} (Tm + Cm) + \sum_{1}^{Nx} (Tc + Cc - Uh) + \sum_{1}^{Ne} (Tm - Tc + Cm - Cc) \]

- **Nm**: Number of compounds not assayed by Eve
- **Tm**: Cost of the time to screen a compound using the mass screening assay
- **Cm**: Cost of the loss of a compound in the mass screening assay
- **Nx**: Number of hits missed by Eve
- **Tc**: Cost of the time to screen a compound using a cherry-picking (confirmation or intelligent) assay
- **Cc**: Cost of the loss of a compound in a cherry-picking assay
- **Uh**: Utility of a hit
- **Ne**: Number of compounds assayed by Eve
Triclosan
Triclosan Repositioned for Malaria

- Simple compound
- Known to be safe – used in toothpaste.
- Targets both DHFR and FAS-II – well established targets.
- Demonstrated activity using multiple wet experimental techniques.
- Works against wild-type and drug-resistant *Plasmodium falciparum*, and *Plasmodium vivax*.
Formalising Science
Formalisation of Science

The goal of science is to increase our knowledge of the natural world through the performance of experiments.

This knowledge should be expressed in formal logical languages.

Formal languages promote semantic clarity, which in turn supports the free exchange of scientific knowledge and simplifies scientific reasoning.
Robot Scientist & Formalisation

- Robot Scientists provide excellent test-beds for the development of methodologies for formalising science.

- Using them it is possible to completely capture and digitally curate all aspects of the scientific process.

- The ontology LABORS is designed to enable the open access of the Robot Scientist experimental data and metadata to the scientific community.

Soldatova et al. (2006) Bioinformatics
Adam’s Investigations

- This formalisation involves >10,000 different research units in a nested tree-like structure 11 levels deep.
- It logically connects >6.6 million OD600\text{_{nm}} measurements to hypotheses, experimental goals, results, etc.
- No previous large-scale experimental work has been so comprehensively described and recorded.
Levels in the Formalisation

Investigation into the automation of Science

Investigation into the automation of novel science

Investigation into the automated discovery of genes encoding orphan enzymes

Automated study of E.C.2.6.1.39 encoding Cycle 1 of automated study of YER152C function

YER152C and Lysine automated trial Experiment 1 (wild-type no metabolite)

Replicate 1 (well) Observation 1
automated study: automated study of yer152c function
has domain of study: functional genomics
has investigator = robot scientist Adam
has goal: 'To test the hypothesis that gene YER152C encodes an enzyme with enzyme class E.C.2.6.1.39'.
has organism of study: Saccharomyces Cerevisiae
has ncbi taxonomy ID: 4932
has hypotheses-set:
has research hypothesis 1: encodes(yer152c,ec_2_6_1_39)
has negative hypothesis 2: not encodes(yer152c,ec_2_6_1_39)
has cycle 1 of study
has study result: the strength of evidence that encodes(yer152c,ec_2_6_1_39):
highest accuracy of random forest evidence: 74%
proportion of random forest evidence >=70%: 2/3
has study conclusion: hypothesis 1 confirmed
Future Prospects
The Future?

- In chess there is a continuum of ability from novices up to Grandmasters.

- I argue that this is also true in science, from the simple research of Adam/Eve, through what most human scientists can achieve, up to the ability of a Newton or Einstein.

- If you accept this, then just as in chess, it is likely that advances in computer hardware and software will drive the development of ever smarter Robot Scientists.

- In favour of this argument are the ongoing development of AI and laboratory robotics.
Vision

- The collaboration between Human and Robot Scientists will produce better science than either can alone – human/computer teams still play better chess than either alone.

- Scientific knowledge will be primarily expressed in logic with associated probabilities and published using the Semantic Web.

- The improved productivity of science leads to societal benefits: better food security, better medicines, etc.
In a 100 years?

- The Physics Nobel Frank Wilczek is on record as saying that in 100 years’ time the best physicist will be a machine.

- A key point about Robot Scientists is that their abilities can be objectively tested.

- So in a 100 years we will know if there are Robot Scientists doing world-class research or not.

- Time will tell.
Conclusions

- Automation is becoming increasingly important in scientific research e.g. DNA sequencing, drug design.

- The Robot Scientist concept represents the logical next step in scientific automation.

- The Robot Scientist Adam is the first machine to have discovered novel scientific knowledge.

- The Robot Scientist Eve has found lead compounds against tropical diseases.
Acknowledgments


- Chicago Big Mechanism consortium.