The Automation of Science
Ross D. King, et al.
Science 324, 85 (2009);
DOI: 10.1126/science.1165620

The following resources related to this article are available online at www.sciencemag.org (this information is current as of April 2, 2009):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:
http://www.sciencemag.org/cgi/content/full/324/5923/85

Supporting Online Material can be found at:
http://www.sciencemag.org/cgi/content/full/324/5923/85/DC1

This article cites 16 articles, 5 of which can be accessed for free:
http://www.sciencemag.org/cgi/content/full/324/5923/85#otherarticles

This article appears in the following subject collections:
Computers, Mathematics
http://www.sciencemag.org/cgi/collection/comp_math

Information about obtaining reprints of this article or about obtaining permission to reproduce this article in whole or in part can be found at:
http://www.sciencemag.org/about/permissions.dtl
The Automation of Science

Ross D. King,1* Jem Rowland,1 Stephen G. Oliver,2 Michael Young,1 Wayne Aubrey,1 Emma Byrne,1 Maria Liakata,1 Magdalena Markham,1 Pinar Pir,2 Larisa N. Soldatova,1 Andrew Sparkes,1 Kenneth E. Whelan,1 Amanda Clare1

The basis of science is the hypothetico-deductive method and the recording of experiments in sufficient detail to enable reproducibility. We report the development of Robot Scientist “Adam,” which advances the automation of both. Adam has autonomously generated functional genomics hypotheses about the yeast Saccharomyces cerevisiae and experimentally tested these hypotheses by using laboratory automation. We have confirmed Adam’s conclusions through manual experiments. To describe Adam’s research, we have developed an ontology and logical language. The resulting formalization involves over 10,000 different research units in a nested treelike structure, 10 levels deep, that relates the 6.6 million biomass measurements to their logical description. This formalization describes how a machine contributed to scientific knowledge.

Computers are playing an ever-greater role in the scientific process (1). Their use to control the execution of experiments contributes to a vast expansion in the production of scientific data (2). This growth in scientific data, in turn, requires the increased use of computers for analysis and modeling. The use of computers is also changing the way that science is described and reported. Scientific knowledge is best expressed in formal logical languages (3). Only formal languages provide sufficient semantic clarity to ensure reproducibility and the free exchange of scientific knowledge. Despite the advantages of logic, most scientific knowledge is expressed only in natural languages. This is now changing through developments such as the Semantic Web (4) and ontologies (5).

A natural extension of the trend to ever-greater computer involvement in science is the concept of a robot scientist (6). This is a physically implemented laboratory automation system that exploits techniques from the field of artificial intelligence (7–9) to execute cycles of scientific experimentation. A robot scientist automatically generates hypotheses to explain observations, devises experiments to test these hypotheses, physically runs the experiments by using laboratory robotics, interprets the results, and then repeats the cycle.

High-throughput laboratory automation is transforming biology and revealing vast amounts of new scientific knowledge (10). Nevertheless, existing high-throughput methods are currently inadequate for areas such as systems biology. This is because, even though very large numbers of experiments can be executed, each individual experiment cannot be designed to test a hypothesis about a model. Robot scientists have the potential to overcome this fundamental limitation.

The complexity of biological systems necessitates the recording of experimental metadata in as much detail as possible. Acquiring these metadata has often proved problematic. With robot scientists, comprehensive metadata are produced as a natural by-product of the way they work. Because the experiments are conceived and executed automatically by computer, it is possible to completely capture and digitally curate all aspects of the scientific process (11, 12).

To demonstrate that the robot scientist methodology can be both automated and be made effective enough to contribute to scientific knowledge, we have developed Robot Scientist “Adam” (13) (Fig. 1). Adam’s hardware is fully automated such that it only requires a technician to periodically add laboratory consumables and to remove waste. It is designed to automate the high-throughput execution of individually designed microbial batch growth experiments in microtiter plates (14). Adam measures growth curves (phenotypes) of selected microbial strains (genotypes) growing in defined media (environments). Growth of cell cultures can be easily measured in high-throughput, and growth curves are sensitive to changes in genotype and environment.

We applied Adam to the identification of genes encoding orphan enzymes in Saccharomyces cerevisiae: enzymes catalyzing biochemical reactions thought to occur in yeast, but for which the encoding gene(s) are not known (15). To set up Adam for this application required (i) a comprehensive logical model encoding knowledge of S. cerevisiae metabolism (~1200 open

References and Notes
17. P. Jäckel, T. Mullin, in Evolutionary Computation in Nanoscale Systems, a U.S. NSF graduate research fellowship, and NSF Creative-IT grant 0757478 and CAREER grant 0547376. We thank M. Kunrnan for editorial consultation and substantive editing of the manuscript.
Adam formulated and tested 20 hypotheses concerning genes encoding 13 orphan enzymes (16) (Table 1). The weight of the experimental evidence for the hypotheses varied (based on observations of differential growth), but 12 hypotheses with no previous evidence were confirmed with $P < 0.05$ for the null hypothesis.

Because Adam's experimental evidence for its conclusions is indirect, we tested Adam's conclusions with more direct experimental methods. The enzyme 2-aminoadipate:2-oxoglutarate aminotransferase (2A2OA) catalyzes a reaction in the lysine biosynthetic pathways of fungi. Adam hypothesized that three genes (YER152C, YJL060W, and YGL202W) encode this enzyme and observed results consistent with all three hypotheses (Table 1). To test Adam's conclusions, we purified the protein products of these genes and used them in vitro enzyme assays, which confirmed Adam's conclusions [supporting online material (SOM)] (Fig. 2).

To further test Adam's conclusions, we examined the scientific literature on the 20 genes investigated (Table 1) (16). This revealed the existence of strong empirical evidence for the correctness of six of the hypotheses; that is, the enzymes were not actually orphans (Table 1). The reason that Adam considered them to be orphans was due to the use of an incomplete bioinformatic database. These six genes therefore constitute a positive control for Adam's methodology. A possible error was also revealed (Table 1) (SOM).

To better understand the reasons why the identity of the genes encoding these enzymes has remained obscure for so long, we investigated their comparative genomics in detail (16). The likely explanation is a combination of three complicating factors: gene duplications with retention of overlapping function, enzymes that catalyze more than one related reaction, and existing functional annotations. Adam's systematic bioinformatic and quantitative phenotypic analyses were required to unravel this web of functionality.

Use of a robot scientist enables all aspects of a scientific investigation to be formalized in logic. For the core organization of this formalization, we used the ontology of scientific experiments: EXPO (11, 12). This ontology formalizes generic knowledge about experiments. For Adam, we developed LABORS, a customized version of EXPO, expressed in the description logic language OWL-DL (17). Application of LABORS produces experimental descriptions in the logic-
programming language Datalog (18). In the course of its investigations, Adam observed 6,657,024 optical density (ODs650nm) measurements (forming 26,495 growth curves). These data are held in a MySQL relational database. Use of LABORS resulted in a formalization of the scientific argument involving over 10,000 different research units (segments of experimental research). This has a nested treelike structure, 10 levels deep, that logically connects the experimental observations to the experimental metadata. (Fig. 3). This structure resembles the trace of a computer program and takes up 366 Mbytes (16). Making such experimental structures explicit renders scientific research more comprehensible, reproducible, and reusable. This paper may be considered as simply the human-friendly summary of the formalization.

Table 1. The orphan enzymes and Adam’s hypotheses. The hypothesized genes are those which Adam abduced encoded an orphan enzyme. Prob. is Adam’s Monte Carlo estimate of the probability of obtaining the observed discrimination accuracy or better with a random labeling of replicates. The discrimination is between the differences in growth curves observed with the addition of specified metabolites to the wild type and the deletant. Acc. is the highest accuracy for a metabolite species in discriminating between the growth curves observed with the addition of specified metabolites to the wild type and the deletant. No. is the number of metabolites tested. Existing annotation is the summary from the Saccharomyces Genome Database of the annotation of the ORF. Dry is the summary of whether the annotated function is the same as predicted by Adam. If a gene already has an associated function, we do not consider this to be contradictory to Adam’s conclusions unless this function is capable of explaining the observed growth phenotype, for example, BCY1. ida indicates inferred from direct assay and iss, inferred from sequence or structural similarity (5). Wet is the result of our manual enzyme assays. See (16) for details.

<table>
<thead>
<tr>
<th>Orphan enzyme</th>
<th>Hypothesized 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 (3.5.99.6)</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>YLI033C (BCY1)</td>
<td>&lt;10⁻⁴</td>
<td>92</td>
<td>11</td>
<td>Cyclic adenosine 3',5'- monophosphate (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 phosphoribosyltransferase (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>Glucosamine-6-phosphate deaminase (3.5.99.6)</td>
<td>YNR034W (SOL3)</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 (THI2)</td>
<td>&lt;10⁻⁴</td>
<td>78</td>
<td>1</td>
<td>Phosphomethylpyrimidine kinase, iss</td>
<td></td>
<td>–</td>
</tr>
<tr>
<td>Mannitol-1-phosphate 5-dehydrogenase (1.1.1.17)</td>
<td>YNR073C</td>
<td>&lt;10⁻⁴</td>
<td>78</td>
<td>6</td>
<td>Putative mannitol dehydrogenase, iss</td>
<td></td>
<td>–</td>
</tr>
<tr>
<td>1-Acylglycerol-3-phosphate O-acyltransferase (2.3.1.51)</td>
<td>YDL052C (SCL1)</td>
<td>0.0001</td>
<td>80</td>
<td>6</td>
<td>1-Acylglycerol-3-phosphate O-acyltransferase ida</td>
<td>✓</td>
<td>–</td>
</tr>
<tr>
<td>Glucosamine-6-phosphate deaminase (3.5.99.6)</td>
<td>YGR248W (SOL3)</td>
<td>0.0002</td>
<td>78</td>
<td>2</td>
<td>6-Phosphogluconolactonase, ida</td>
<td></td>
<td>–</td>
</tr>
<tr>
<td>Maleylacetacetoacetyl isomerase (5.2.1.2)</td>
<td>YLL060C (GTT2)</td>
<td>0.0003</td>
<td>76</td>
<td>3</td>
<td>Glutathione S-transferase, ida</td>
<td></td>
<td>–</td>
</tr>
<tr>
<td>Serine O-acetyltransferase (2.3.1.30)</td>
<td>YLR218W</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>YJR060W (BA3)</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, iss</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>YJR045W</td>
<td>0.1300</td>
<td>72</td>
<td>1</td>
<td>Succinate dehydrogenase isozyme, iss</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>
Fig. 2. Assay results for 2A2OA activity. The proteins encoded by YGL202W, YJL060W, YER152C, and YDL168W were expressed from OpenBiosystems (www.openbiosystems.com) yeast ORF clones and purified. Activity was tested in an assay of NADPH (reduced form of nicotinamide adenine dinucleotide phosphate) production based on (22). L-ε-aminoadipic acid and 2-oxoglutarate were provided as substrates and pyridoxal phosphate as cofactor. Glutamate production was assayed by using commercially available yeast glutamate dehydrogenase, which uses NADP as cofactor and deaminates yeast glutamate dehydrogenase, which uses NADP as cofactor and deaminates glutamate, producing ammonia and NADPH and regenerating 2-oxoglutarate (16). Also consistent with 2A2OA activity is experimental evidence indicating a higher activity with L-glutamate, producing ammonia and NADPH and regenerating 2-oxoglutarate (uses NADP as cofactor and deaminates yeast glutamate dehydrogenase, which uses NADP as cofactor and deaminates glutamate, producing ammonia and NADPH and regenerating 2-oxoglutarate (16)).

Fig. 3. Structure of the Robot Scientist investigation (a fragment). It consists of two main parts: an investigation into the automation of science and an investigation into the reuse of formalized experiment information. The top levels involve AI research (red), which requires research in functional genomics (blue) and systems biology (yellow). Each level of research unit (studies, cycles, trials, tests, and replicates) is characterized by a specific set of properties (fig. S3) (16). Such a nested structure is typical of many scientific experiments, where the testing of a top-level hypothesis requires the planning of many levels of supporting work. What is atypical in Adam’s work is the scale and depth of the nesting.

A major motivation for the formalization of experimental knowledge is the expectation that such knowledge is more easily reused to answer other scientific questions. To test this, we investigated whether we could reuse Adam’s functional genomic research (16). An example question investigated was the relative growth rates ($\mu_{max}$) in rich and defined media of the deletion strains compared with those of the wild type. What was observed, in both media, was a skewed distribution, with a few deletants having a much lower $\mu_{max}$ than that of the wild type, but most having a slightly higher $\mu_{max}$. These observations question the common assumption that wild-type *S. cerevisiae* is optimized for $\mu_{max}$ and provide quantitative test data for yeast systems biology models (19).

It could be argued that the scientific knowledge “discovered” by Adam is implicit in the formulation of the problem and is therefore not novel. This argument that computers cannot originate anything is known as Lady Lovelace’s objection (20). “The Analytical Engine has no pretensions to originate anything. It can do whatever we know how to order it to perform” (her italics). We accept that the knowledge automatically generated by Adam is of a modest kind. However, this knowledge is not trivial, and in the case of the genes encoding 2A2OA, it sheds light on, and perhaps solves, a 50-year-old puzzle (21).

Adam is a prototype and could be greatly improved. Its hardware and software are “brittle,” so although Adam is capable of running for a few days without human intervention, it is advisable to have a technician nearby in case of problems. The integration of Adam’s artificial intelligence (AI) software also needs to be enhanced so that it works seamlessly. To extend Adam, we have developed software to enable external users to propose hypotheses and experiments, and we plan to automatically publish the logical descriptions of automated experiments. The idea is to develop a way of enabling teams of human and robot scientists to work together. The greatest research challenge will be to improve the scientific intelligence of the software. We have shown that a simple form of hypothesis-led discovery can be automated. What remain to be determined are the limits of automation.

References and Notes
9. 3 APRIL 2009 VOL 324 SCIENCE www.sciencemag.org

88.
Priming in Systemic Plant Immunity

Ho Won Jung,1 Timothy J. Tschaplinski,2 Lin Wang,3* Jane Glazebrook,3 Jean T. Greenberg1†

Plants possess inducible systemic defense responses when locally infected by pathogens. Bacterial infection results in the increased accumulation of the mobile metabolite azelaic acid, a nine-carbon dicarboxylic acid, in the vascular sap of Arabidopsis thaliana. Internal infection becomes primed to previously unexposed (naïve) plants (1). Leaves infected with SAR-inducing bacteria produce vascular sap, called petiole exudate, which confers disease resistance to previously unexposed (naïve) plants (2, 3). This indicates that a mobile systemic signal(s) is involved in SAR (4). Although the hormone jasmonic acid (JA) accumulates to a high level in petiole exudates from leaves infected with SAR-inducing bacteria, JA does not seem to be the critical signal for SAR (5, 6). Instead, SAR and the production of active exudates require the DIR1 protein, a predicted secreted protein and putative signal carrier in the lipid transfer protein family, and other proteins involved in glycerolipid biosynthesis (2, 3, 7). Additionally, SAR and exudate-induced resistance appears to require the phenolic metabolite salicylic acid (SA) (3, 8) and possibly methylsalicylate (MeSA) and its methyl ester.

Whole plant immunity, called systemic acquired resistance (SAR), often develops after localized foliar infections by diverse pathogens. In this process, leaves distal to the localized infection become primed to activate a stronger defense response upon secondary infection (1). Leaves infected with SAR-inducing bacteria produce vascular sap, called petiole exudate, which confers disease resistance to previously unexposed (naïve) plants (2, 3). This indicates that a mobile systemic signal(s) is involved in SAR (4). Although the hormone jasmonic acid (JA) accumulates to a high level in petiole exudates from leaves infected with SAR-inducing bacteria, JA does not seem to be the critical signal for SAR (5, 6). Instead, SAR and the production of active exudates require the DIR1 protein, a predicted secreted protein and putative signal carrier in the lipid transfer protein family, and other proteins involved in glycerolipid biosynthesis (2, 3, 7). Additionally, SAR and exudate-induced resistance appears to require the phenolic metabolite salicylic acid (SA) (3, 8) and possibly methylsalicylate (MeSA) and its methyl ester.

Materials and methods are available as supporting material on Science Online.

8. Materials and methods are available as supporting material on Science Online.
15. This work was funded by grants from the Biotechnology and Biological Sciences Research Council to R.D.K., S.G.D., and M.Y.; by a SRF 2 award to R.D.K.; by fellowships from the Royal Commission for the Great Exhibition of 1851, the Engineering and Physical Sciences Research Council, and the Royal Academy of Engineering to A.C.; and by a RCUK Fellowship to L.N.S. We thank D. Struttman for help with yeast and M. Benway for help with Adam.

Supporting Online Material
www.sciencemag.org/cgi/content/full/324/5923/85/DC1
Materials and Methods
Figs. S1 to S3
Table S1
References
8 September 2008; accepted 10 February 2009
10.1126/science.1165620

Fig. 1. Azelaic acid specifically confers resistance to Pseudomonas syringae. (A) Azelaic acid–induced resistance is concentration-dependent. Plants were sprayed with 1, 10, 100, and 1000 μM azelaic acid in 5 mM MES (pH 5.6) or 5 mM MES (pH 5.6) alone 2 days before infection with P. syringae pv. maculicola strain PmoDG3 (OD600 = 0.0001). (B) Induced resistance is time-dependent. Plants sprayed with 5 mM MES or 1 mM azelaic acid for the time periods indicated were subsequently inoculated with PmoDG3. (C) 5 mM MES or 1 mM azelaic acid was injected into local leaves. Two days later, either local or systemic leaves were infected with PmoDG3. (D) Dicarboxylic acids (1 mM) of different carbon-chain lengths were applied to Arabidopsis. M, 5 mM Mes; C9, suberic acid; C8, azelaic acid; C10, sebacic acid. (E) Mobility of deuterium-labeled azelaic acid [HOOC(CD2)7COOH] injected into WT leaves. Azelaic acid amounts were determined in petiole exudates (left) and distal leaves (right) after local injection with 1 mM azelaic acid. ***P < 0.05; **P < 0.01; t test. Error bars indicate SE.