Supporting Online Material

1. Introduction

This file contains a description of the most important Supporting Online Material (SOM). A full catalogue of the SOM can be found at http://www.aber.ac.uk/compsci/Research/bio/robotsci/data/.

The largest piece of SOM is the annotation of Adam's experiments using LABORS. This contains details of the over 10,000 different research-units, and takes up 366 Megabytes of Datalog. This annotation is designed to be analyzed by computer. Further information: See the data/ subdirectory.

2. Materials and methods

2.1 Logical model encoding knowledge of *S. cerevisiae* metabolism

The full details of the model are given in Whelan & King (*15*). The model is in essence a directed labelled hyper-graph (with metabolites as nodes and enzymes as arcs) encoded in Prolog (*s1*). This model has been shown to be as good as state-of-the-art Flux Balance Analysis models at predicting whether gene deletions are lethal (*15*).

Further information:

model/index.html

2.2 Bioinformatic Database

The core bioinformatic database we used was mostly taken from KEGG (*s2*). We fixed the database in October 2006.

Further information: informatics/index.html

2.3 Software to abduce hypotheses

In our previous work (6) we used a purely logical approach to hypothesis generation. This method is too inefficient for a model of the size of yeast metabolism. We therefore developed an alternative approach based on bioinformatic methods. This is based on the insight that one way of considering standard genome annotation is as a vast process of abductive hypothesis formation (*s3*). Adam's hypothesis generation method was as follows: *i*. Find all reactions in its model of *S.*

cerevisiae metabolism which are orphans. *ii*. Compute (deduce) which of these, when removed, are expected to interfere with cell growth - there is no longer a path from start to end-point metabolites (*15*). *iii*. Find the EC enzyme class of these reactions. *iv*. Find genes from other organisms that code for enzymes of the same EC class. *v*. Use sequence similarity searches to identify the most likely homologs to those genes in *S. cerevisiae*. *vi*. Hypothesize (abduce) that these identified genes code for the orphan enzyme. In evolutionary/logical terms: a common ancestral gene and descent with conservation of function are abduced, and these explain the observed conservation of sequence.

Further information: procedure/procedure_infer_hyps_relating_enzymes_and_genes_novel.txt

2.4 Software to deduce experiments

To select experiments to test hypotheses Adam uses its logical model of metabolism to identify metabolites within two metabolic reactions of the orphan enzyme. Adam then compares these metabolites with its list of available metabolites.

If a hypothesis is correct then the expected observation is: that the difference in growth between the deletion strain growing on defined medium and on defined medium + metabolite can be discriminated from the difference in growth between the wild type growing on defined medium and defined medium + metabolite.

Further information: procedure/procedure_derive_hyps_relating_metabolites_and_genes.txt

2.5 Software to plan and design the experiments

To plan experiments, Adam uses classical experimental design techniques. In the research described here a two-factor design was used: wild type v deletion strain, defined medium v defined medium + metabolite; 96-well microtitre plates were used giving 24 repeats of each strain/medium combination. To control for intra-plate environmental effects, six Latin-squares were used per plate. Manual use of such experimental design techniques using 96-well microtitre plates is impractical.

Further information: procedure/procedure_to_design_experiment_plates/

2.6 Laboratory automation software

The software to control the laboratory automation of Adam was written by Caliper Life Sciences (closed source).

2.7 Software to generate growth curves and extract parameters

Adam's observations are discrete OD_{595nm} measurements. These are stored in a relational database. The first step in analyzing these is to use spline fitting to form continuous growth curves for each replicate (one well). From these growth curves biologically relevant parameters are extracted such as lag-time, μ_{max}, and maximum cell density.

Further information:

procedure/procedure_to_calculate_pregrowth_and_growth_curve_parameters/index.html

2.8 Software to relate the analyzed data to the hypotheses

Most deletion mutants of *S. cerevisiae* that exhibit clear qualitative phenotypes, such as auxotrophy, have already been discovered. This means that the remaining deletion mutants are expected to have only quantitative differences in growth compared to the wild type. Therefore, statistical methods are required to identify significant differences.

We used decision trees and forests, with resampling methods, to decide whether the difference in growth between the deletion strain growing on defined medium and on defined medium + metabolite could be discriminated from the difference in growth between the wild type growing on defined medium and defined medium + metabolite. We checked these results using classical statistical methods.

Further information:

procedure/procedure_to_calculate_stats_of_growth_curve_parameters.html,

procedure/procedure_to_automatically_interpret_trial_results_novel.html,

procedure/procedure_strength_of_evidence_integrated_trials.html

3. Full Automation of a Robot Scientist

To demonstrate the automation of cycles of experiment using a Robot Scientist we programmed Adam to repeat the experiments of our first, semi-automated, Robot Scientist (*6*). These experiments concern the rediscovery of functional genomics knowledge about the aromatic amino acid biosynthesis pathway in *S. cerevisiae*. This experimental problem differs subtly from that in the main text: given a known correct model of aromatic amino acid metabolism identify the enzyme(s) encoded by deleted genes (*6*). The comparison between the previous gene-function prediction experiments and those performed by Adam is shown in figs S1 and S2. The results for Adam are slightly better than the original. This demonstrates that cycles of experiment can be automated by a Robot Scientist. It also confirms the first Robot Scientist's results.

Further information: index.html, section “Investigation into full automation of AAA experiments”

4. Experimental confirmation of Adam's conclusions

4.1. Background

The enzyme class 2-aminoadipate:2-oxoglutarate aminotransferase (2A2OA) catalyses the reaction: α-ketoadipic acid + L-glutamate ↔ L-α-aminoadipic acid + α-oxoglutarate (E.C.2.6.1.39), a step in the lysine biosynthetic pathways of fungi. The enzyme has been studied since the 1950s because of its role in antibiotic production, and also as a possible antibiotic target because fungi use a different lysine synthesis pathway to that used by animals (*17*). Despite this long-term attention, the identity of the gene(s) coding for the enzyme 2A2OA is still unclear - research on transaminases is complicated because they are often promiscuous in their use of substrates. In the 1960s, it was found that yeast has at least two iso-enzymes for 2A2OA (*s4*).

4.2 Bioinformatics

KEGG was the only bioinformatic source we found with an annotation for this enzyme: it identifies YGL202W (*ARO8*) as encoding 2A2OA. We believe this annotation comes from Urrestarazu and colleagues (*s5*) who demonstrated that the gene product of *ARO8* has 2A2OA activity and state that it “might be identical to one of the two known alpha-aminoadipate aminotransferases”.

4.3 Adam's study

In Adam's bioinformatic database there was only one gene encoding a product with the EC annotation 2.6.1.39 (AadAT from *R. norvegicus*). Adam used the sequence similarity search method FASTA3.4 to identify three possible yeast homologs to AadAT: YER152C, YJL060W, & YGL202W (in order of similarity). Adam therefore abduced three hypotheses: YER152C encodes 2A2OA, YJL060W encodes 2A2OA, and YGL202W encodes 2A2OA. Thus strains deleted for either YER152C, YJL060W, or YGL202W will grow differently from the wild type on media supplemented by the addition of metabolites involved in the lysine pathway near 2A2OA. Three metabolites were available (L-lysine, L-2-aminoadipate, L-saccharopine). Adam executed these growth experiments and observed results consistent with all three hypotheses.

4.4. Manual study

Full details of the enzyme assays can be found here: manual_bio_study/index.html

To further test the hypotheses, we manually constructed all three possible double deletion strains: a ygl202wΔ ygl060Δ double mutation is lethal, the difference in phenotype between the ygl202wΔ yer152cΔ double mutant grown on defined medium and on defined medium + metabolite is greater than for either of the single deletants on their own, the phenotype of the ygl060wΔ yer152cΔ double mutant is similar to that of a ygl060wΔ single deletant. These results are consistent with Adam's conclusions.

Further information:

data/study_YER152C_YGL202W/

data/study_YER152C_YGL202W.pl

data/study_YER152C_YJL060W/

data/study_YER152C_YJL060W.pl

Also see the PhD thesis of Wayne Aubrey ("Towards the Automation of Selection and Construction of Multiple Gene Deletant Strains of *Saccharomyces cerevisiae*" 2009, Aberystwyth University).

The recent work of Wogulis *et al.* (*s6*) has demonstrated that YJL060W has kynurenine aminotransferase activity, and is not an arylformamidase as previously annotated. This is significant because enzymes with both kynurenine aminotransferase and 2A2OA activity are known in yeast and other organisms (*17*).

4.5. Conclusions

Integrating these different stands of evidence we conclude that Adam's conclusions were correct: YGL202W is correctly annotated by KEGG as encoding a 2A2OA; YJL060W (the open reading-frame currently annotated as encoding a kyneurenine aminotransferase) encodes a 2A2OA, and YER152C (currently not annotated with any function) also probably encodes a 2A2OA.

5. Literature analysis of Adam's conclusions

Examination of the biological literature revealed good empirical evidence for the validity of six of the hypotheses. These six genes therefore constitute a positive control for Adam's methodology: Adam correctly hypothesized the gene encoding an enzyme and found evidence consistent with the hypothesis. Note that the concentration on orphan enzymes in **2.3** is simply a heuristic to focus Adam on investigating areas of metabolism where new scientific knowledge is most likely to be found. In the case of the reaction catalyzed by 2A2OA, which was not an orphan (as it is annotated with the gene YGL202W - at least in KEGG), Adam still managed to correctly identify YJL060W (and possibly YER152C) as also encoding 2A2OAs. This illustrates one possible role for Robot Scientists: to automatically verify existing bioinformatic annotations using wet experimentation.

We carefully examined the bioinformatic databases and literature on all of the genes examined. We particularly focused on the ida cases.

YLR017W - Adam was correct in its hypothesis, but the experimental evidence it found was weak. It is arguably an error of omission. Note that *PNP1*/YLR209c and *MEU1*/YLR017w are paralogous genes that arose from a local duplication. This example places a floor on the amount of evidence required for a correct hypothesis.

YLR070C - The literature evidence for xylitol dehydrogenase function (*s7*) does not derive from experiments on the purified protein, nor has dual functionality been excluded.

YLL060C - Both Adam's function and the literature's function are plausible since, there are known genes with both functions e.g. hGSTZ1-1 (*s8*).

YGR248W and YHR163W – These are currently annotated as 6-phosphogluconolactonases. Adam's hypothesis of glucosamine-6-phosphate deaminase is closely related both chemically and enzymatically. The Nag family of proteins which normally act as glucosamine-6-phosphate deaminases are absent from *S. cerevisiae,* and the closely related Sol proteins are expanded. The whole-cell assays used in the literature for 6-phosphogluconolactonase do not necessarily discriminate between the two functions (*s9*).

YDL168W - is currently annotated as a multifunctional enzyme possessing both alcohol dehydrogenase and glutathione-dependent formaldehyde dehydrogenase activities that function in formaldehyde detoxification and the formation of long chain and complex alcohols. The gene product is also involved in the degradation of other amino acids. Therefore, given the hydroxyl group of threonine, the existing annotation does not seem contradictory to Adam's conclusion (*s10*).

The manual examination of the literature revealed one likely error of commission by Adam. This is for ORF YIL033C (*BCY1*), which Adam predicted to be a glutaminase (E.C.3.5.1.2). Adam found that a yil033cΔ mutant has a clearly different growth profile from the wild type ($P < 10^{-4}$) (Table 1). All 11 metabolites predicted to have differential effect on a glutaminase deletant affected the yil033cΔ strain. This is robust experimental evidence consistent with Adam's hypothesis. However, there is good experimental evidence that YIL033C has a cAMP-dependent protein kinase regulatory subunit involved in the control of enzymes; and this function may be sufficient to explain the observed phenotypes. This possible mistake exposes a weakness in Adam's current metabolic model, which does not include any control mechanisms. It is however possible that YIL033C is both a kinase and a glutaminase, and it is intriguing that yil033c mutants are known to be sensitive to ammonium starvation and that this sensitivity varies for different point mutants (*s11*).

6. Comparative Genomics

The discovery of the genes encoding these enzymes investigated by Adam is presumably particularly difficult, given that decades of research failed to find them. The reason for this dificulty is not because these are specialist genes that are peculiar to *S. cerevisiae* and its near relatives. Of the 20 ORFs only one (YLL060C) represents a gene that is confined to the *Saccharomyces* 'sensu stricto' clade (*s12, s13*). Of the other 19 ORFs, an MCL cluster analysis, using the e-Fungi database (*s13*) showed that only three (YER183C, YHR163W, and YLR017C) encode products that are not members of any protein families found in other yeast and fungal species. Moreover, ten encode products that are members of families in which at least one other member has a function congruent with that proposed by Adam. It seems most likely that the functions of these genes have remained undiscovered because of the redundancy within the *S. cerevisiae* genome (*s14*). No less than 14 of the 20 ORFs have paralogs elsewhere in the yeast genome (*s15*), with 10 being paralogs of other members of the set of 20 (see Table, below). These gene duplications arose in various ways. Two of the ORFs (YGR248W and YHR163W) form a paralogous pair that was formed by the whole-genome duplication (WGD) that occurred in the evolutionary history of the *Saccharomyces* clade (*s12*), while four others (YDL168W, YGL202W, YGR248W, and YNR027W) probably arose in ancient duplication events that preceded the WGD. Five of the ORFs are telomere-associated (see Table S3) and yeast chromosome ends are well known for their plasticity and genetic redundancy (*s14*). It is clear that classical genetics, and even functional genomics, would find it difficult to uncover the functions of this set of duplicated and inter-related genes. However, their functions have been revealed by the model-driven systematic bioinformatic and quantitative phenotypic analyses performed by Adam.

7. Ontology

Examining Figure 3, a typical path through the LABORS formalization is the following: The Robot Scientist investigation has a part (the investigation into the automaton of science which has a part (the investigation into whether the Robot Scientist Adam can discover novel science which has a part (the study aimed at finding the genes encoding orphan enzymes which has a part (the study of the orphan enzyme E.C.2.6.1.39 in *S. cerevisiae* which has a part (the cycle of studies of the gene

YER152C which has a part (the study of the gene YER152C which has a part (the cycle 1 which has a part (the trial of the compound C00047/lysine which has a part (the test of addition of C00047/lysine which has a part (the replicate 1 which has the parts (298 growth and 51 pre-growth observations)))))))))).

Figure S3 shows an example of one structural unit, the automated study of YER152C function. The metadata are represented in free text, OWL (*19),* and Datalog (*20*) clauses.

Further information: data/LABORS.owl

8. Reuse of data

A major motivation for the formalisation of knowledge is the hypothesis that such knowledge is easier to reuse in other scientific investigations. To test this we investigated whether we could reuse the results of Adam's "Investigation into the automation of novel science" to investigate Systems Biology. These investigations examined the quantitative relationships between genotype, environment, and phenotype. This has previously been investigated in yeast using specialised batch systems (*s16*), and in continuous culture (*s17*). Two investigations were undertaken: an investigation into the relative growth of deletion mutants v wild type in rich and defined media, and an investigation into the relative growth of the wild type with the addition of metabolites to the defined medium.

Further information: data_reuse/

9. Figures and Table

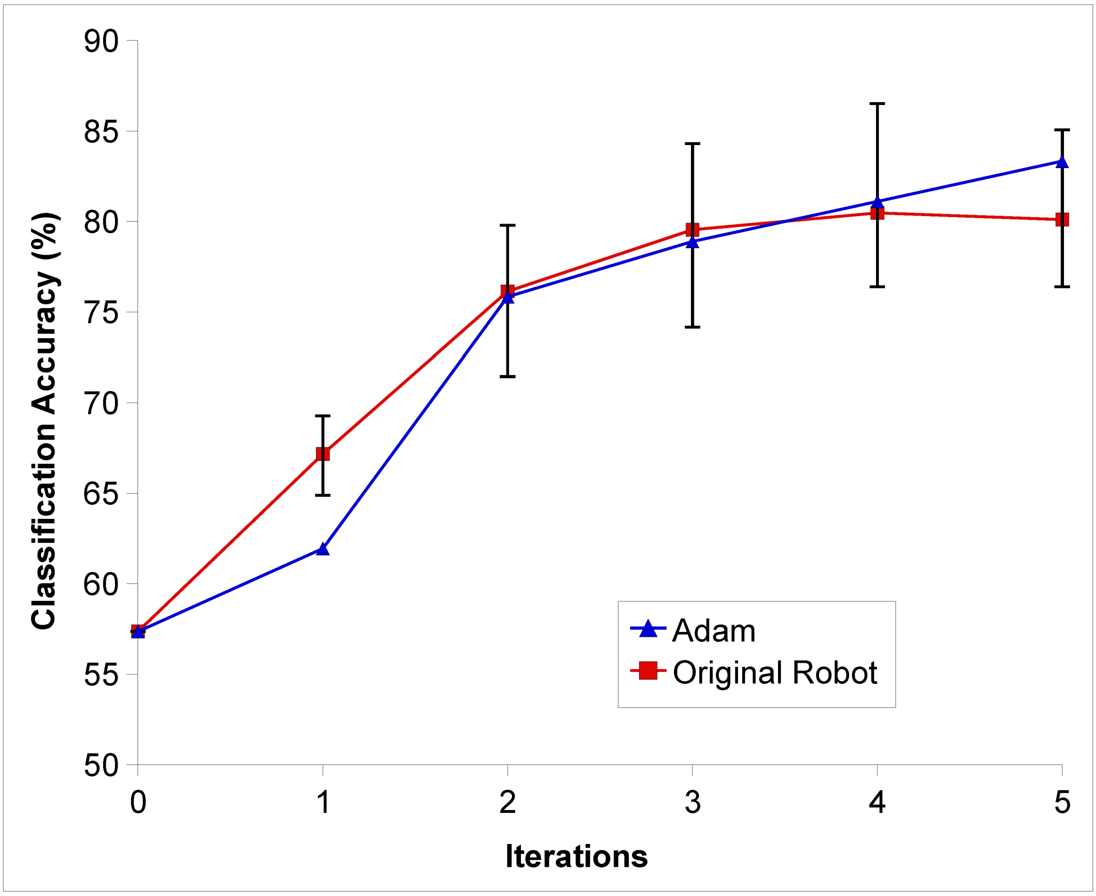


fig S1 Demonstration of cycles of study. Comparison of classification accuracy versus experimental iteration (time) for Adam and the average of the original Robot Scientist on the repeat of the original functional genomics problem. The error bars are the highest and lowest results from the original work.

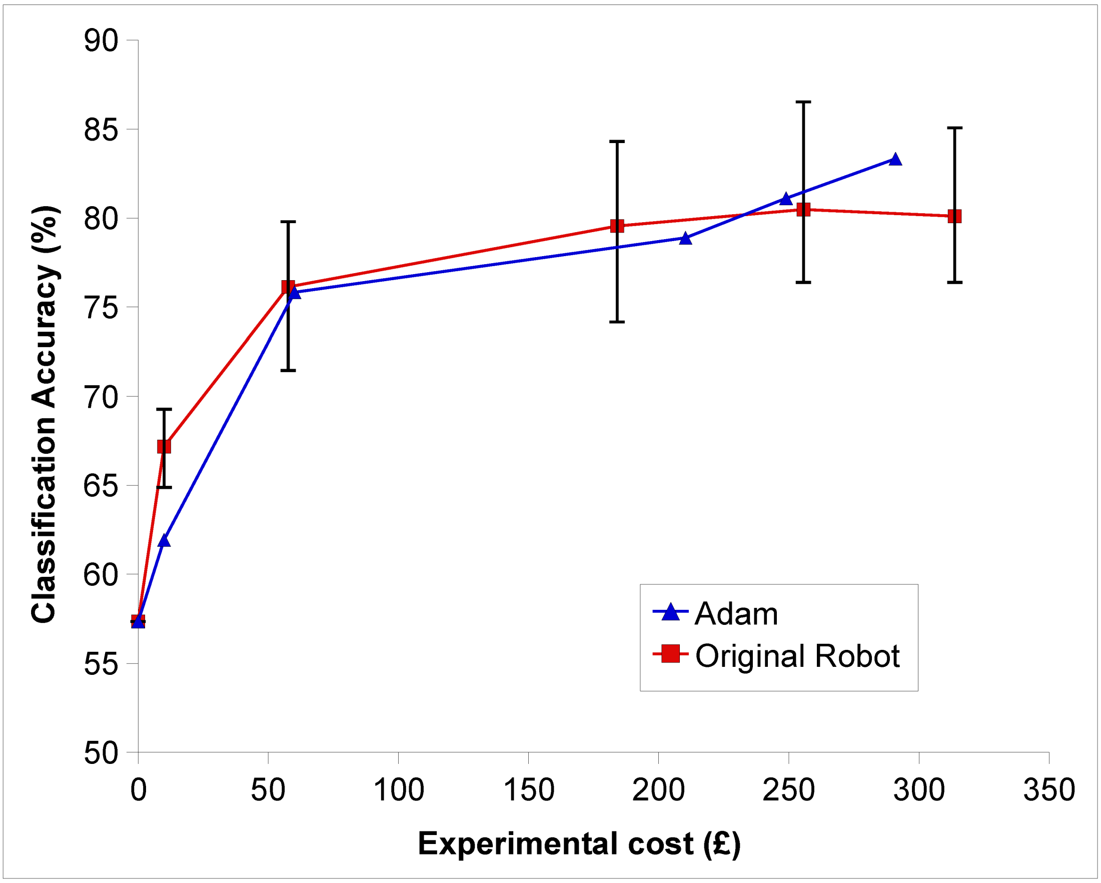
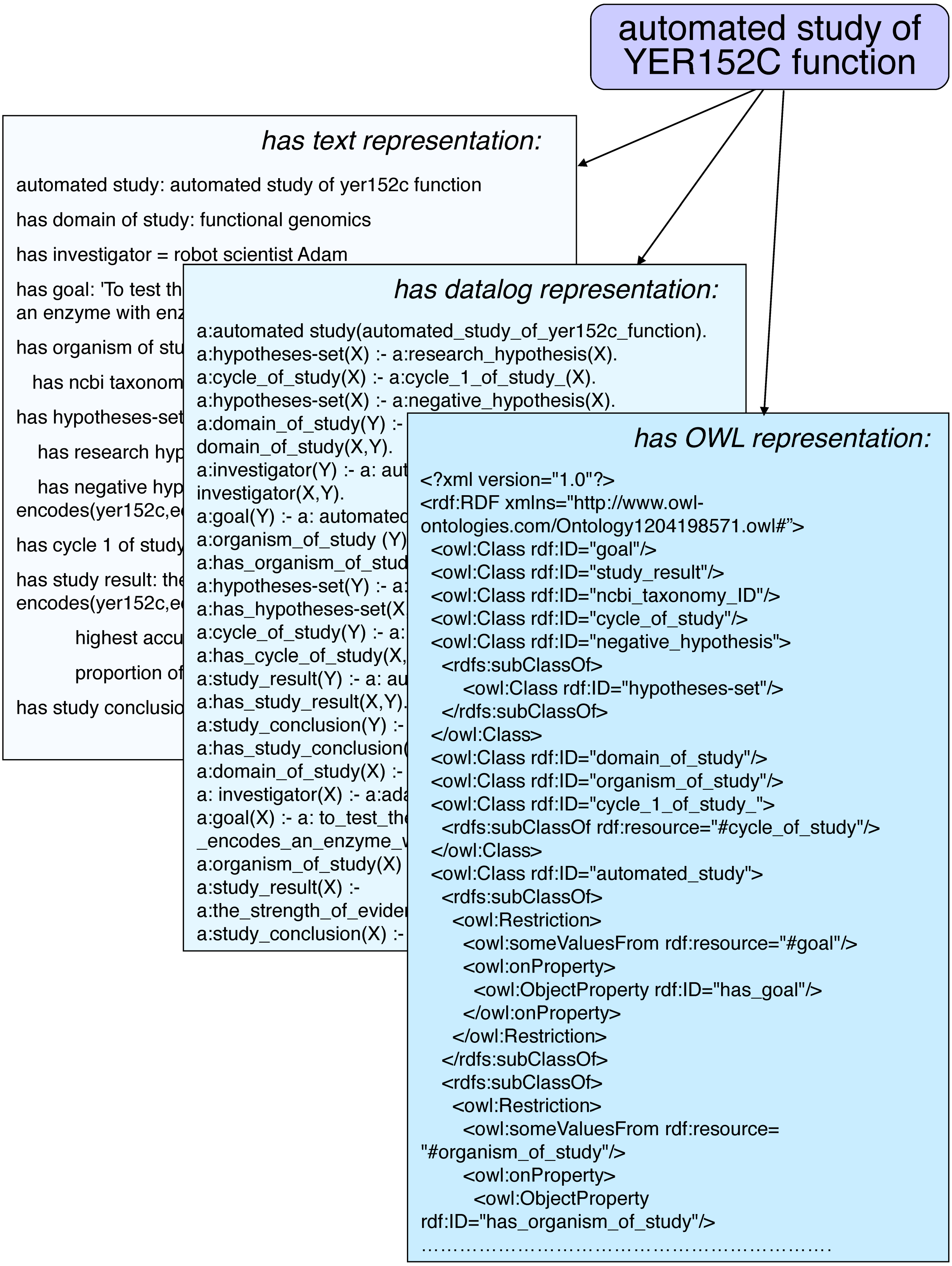


fig S2 Demonstration of cycles of study. Comparison of classification accuracy versus experimental cost (pounds sterling) for Adam and the average of the original Robot Scientist for the repeat of the original functional genomics problem.



has text representation:

automated study: automated study of yer152c function
has domain of study: functional genomics
has investigator = robot scientist Adam
has goal: 'To test th
an enzyme with enz
has organism of stu
has ncbi taxonom
has hypotheses-set
has research hyp
has negative hyp
encodes(yer152c,e
has cycle 1 of study
has study result: the
encodes(yer152c,e
highest accu
proportion of
has study conclusio

has datalog representation:

```
a:automated study(automated_study_of_yer152c_function).
a:hypotheses-set(X) :- a:research_hypothesis(X).
a:cycle_of_study(X) :- a:cycle_1_of_study_(X).
a:hypotheses-set(X) :- a:negative_hypothesis(X).
a:domain_of_study(Y) :-
domain_of_study(X,Y).
a:investigator(Y) :- a: aut
investigator(X,Y).
a:goal(Y) :- a: automatec
a:organism_of_study (Y)
a:has_organism_of_stud
a:hypotheses-set(Y) :- a:
a:has_hypotheses-set(X
a:cycle_of_study(Y) :- a:
a:has_cycle_of_study(X,
a:study_result(Y) :- a: au
a:has_study_result(X,Y).
a:study_conclusion(Y) :-
a:has_study_conclusion(
a:domain_of_study(X) :-
a: investigator(X) :- a:ada
a:goal(X) :- a: to_test_the
_encodes_an_enzyme_v
a:organism_of_study(X)
a:study_result(X) :-
a:the_strength_of_evider
a:study_conclusion(X) :-
```

has OWL representation:

```
<?xml version="1.0"?>
<rdf:RDF xmlns="http://www.owl-
ontologies.com/Ontology1204198571.owl#">
  <owl:Class rdf:ID="goal"/>
  <owl:Class rdf:ID="study_result"/>
  <owl:Class rdf:ID="ncbi_taxonomy_ID"/>
  <owl:Class rdf:ID="cycle_of_study"/>
  <owl:Class rdf:ID="negative_hypothesis">
    <rdfs:subClassOf>
      <owl:Class rdf:ID="hypotheses-set"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="domain_of_study"/>
  <owl:Class rdf:ID="organism_of_study"/>
  <owl:Class rdf:ID="cycle_1_of_study_">
    <rdfs:subClassOf rdf:resource="#cycle_of_study"/>
  </owl:Class>
  <owl:Class rdf:ID="automated_study">
    <rdfs:subClassOf>
      <owl:Restriction>
        <owl:someValuesFrom rdf:resource="#goal"/>
        <owl:onProperty>
          <owl:ObjectProperty rdf:ID="has_goal"/>
        </owl:onProperty>
      </owl:Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
      <owl:Restriction>
        <owl:someValuesFrom rdf:resource=
"#organism_of_study"/>
        <owl:onProperty>
          <owl:ObjectProperty
rdf:ID="has_organism_of_study"/>
…………………………………………………………
```

fig S3 Example of one structural unit, the automated study of YER152C function. The metadata are represented in freetext, OWL (15), and Datalog (16) clauses.

Paralogs and Telomere Association of Genes assigned to Orphan Reactions by Adam

Genes assigned to orphan reactions by Adam	Telomere Associated?	Closest Paralog	Telomere Associated?	% Identity	e-Value	Closest Paralog in the Robot group	Telomere Associated?	% Identity	e-Value
YDL052C									
YDL168W		YBR145W		35%	3.00 e-21	YLR070C		27%	5.90 e-9
YER152C		YGL202W		27%	1.30 e-7	YGL202W		27%	1.30 e-7
YER183C	YES								
YGL202W		YHR137W		30%	9.20 e-53	YER152C		27%	1.30 e-7
YGR248W	YES	YHR163W		48%	2.00 e-66	YHR163W		48%	2.00 e-66
YHR163W		YGR248W	YES	48%	2.00 e-66	YGR248W	YES	48%	2.00 e-66
YIL033C									
YJL045W		YKL148C		81%	1.50 e-290				
YJL060W		YDR111C		25%	2.70 e-7	YGL202W		23%	2.50 e-5
YJL218W	YES								
YLL060C	YES								
YLR017W		YLR209C		33%	2.60 e-6	YLR209C		33%	2.60 e-6
YLR070C		YJR159W	YES	54%	9.90 e-104	YDL168W		30%	7.5 e-15
YLR209C		YLR017W		33%	2.40 e-6	YLR017W		33%	2.4 e-6
YMR020W									
YNR027W		YEL029C		37%	6.2 e-47				
YNR034W		YCR073W-A		74%	1.30 e-124	YHR163W		37%	1.2 e-34
YNR073C	YES	YEL070W	YES	100%	8.9 e-277				
YPR121W		YPL258C	YES	74%	1.1 e-232				

table S1 Paralogs and Telomere Association of Genes assigned to Orphan Reactions by Adam. Paralogs of the genes studied by Robot Scientist Adam were found using default parameters in the online tool WU-BLAST2, the amino acid sequences encoded by the genes were retrieved from SGD (*s15*) as an input to BLASTP search. The closest paralogs in the genome and within Adam's list are listed in the table; paralogs with an e-value higher than 1e-4 were ignored. The complete BLASTP results for the genes are given, for the closest paralogs, the alignment of amino-acid sequences are also given, together with the statistics.

Locations of the genes on the chromosomes were downloaded from SGD (*s15*) and the genes located within 10% of the total length of the chromosome from either chromosome end are designated as 'telomere-associated' in the Table.

11. References

s1. Reiser, P.K., *et al.*, *Electronic Transactions in Artificial Intelligence*. **5**, 233 (2001).

s2. Kanehisa, M., Goto, S. *Nucleic Acids Res.* **28**, 27 (2000).

s3. Pierce, C.S. *Collected Papers of Charles Sanders Pierce*. Harvard University Press, Cambridge, MA, (1958).

s4. Masuda, M., Ogur, M. *J. Biol. Chem.* **244**, 5153 (1969)

s5. Urrestarazu, A. *et al.*, *Mol. Gen. Genet.* **257**, 230 (1998).

s6. Wogulis, M., *et al.*, *Biochemistry.* **47,** 1608 (2008).

s7. Richard, P. *et al.*, *FEBS Lett.* **457**, 135 (1999).

s8. Board, P.G. *et al., Biochem. J.* **374** 731 (2003).

s9. Stanford, D.R. *et al., Genetics*, **168**, 117 (2004).

s10.Dickinson, J.R. *et al.*, *J. Biol. Chem.* **278,** 8028 (2003).

s11.Cannon, J.F., *et al.*, *J. Biol. Chem.* **265,** 11897 (1990).

s12.Byrne, K.P., Wolfe, K.H. *Genome Res.* **15,** 1456 (2005)

s13.Cornell, M.J. *et al. Genome Res.* **11,** 1809 (2007).

s14.Goffeau, A. *et al. Science* **274**, 546-567 (1996).

s15.SGD project. “Saccharomyces Genome Database” http://www.yeastgenome.org/ (2008).

s16.Warringer, J. *et al.*, *Proc. Natl. Acad. Sci USA* **100,** 15724 (2003).

s17.Delneri, D. *et al. Nature Genetics* **40**, 113 (2008).