Refining Metabolic Network Models in the Robot Scientist Project

BMI/CS 776
www.biostat.wisc.edu/bmi776/
Mark Craven
craven@biostat.wisc.edu
Spring 2011

Goals for Lecture

the key concepts to understand are the following

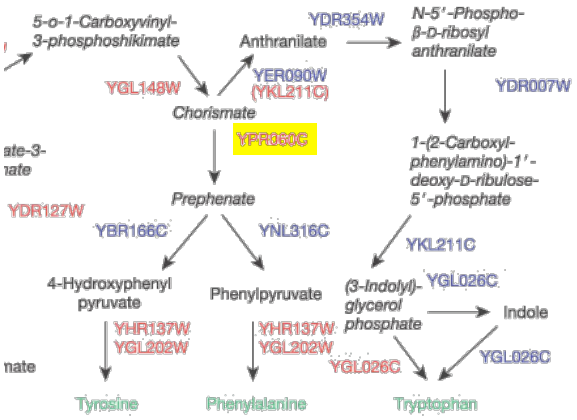
- auxotrophic growth experiments
- the experiment selection task
- closed-loop experimentation

The *Robot Scientist* Project

[King et al., *Nature* 2004; King et al., *Science* 2009]

- developed a closed-loop system that tries to refine a metabolic network model by iteratively
  - determining next experiment to run
  - executing experiment
  - reading result

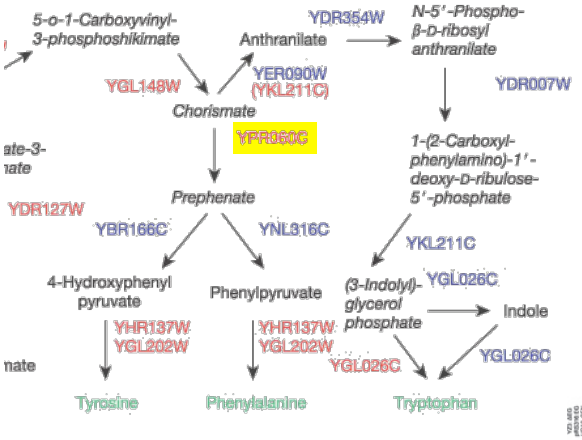
Auxotrophic Growth Experiments



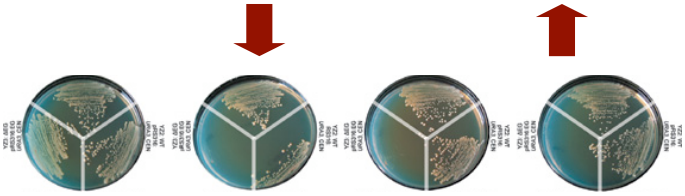
Does YPR060C catalyze the reaction that produces prephanate from chorismate?

- for various combinations of genetic mutants and growth media, determine whether cells were able to grow or not (or measure growth curve across multiple time points)
- a knockout mutant is *auxotrophic* if it cannot grow on a medium on which the wild type can grow

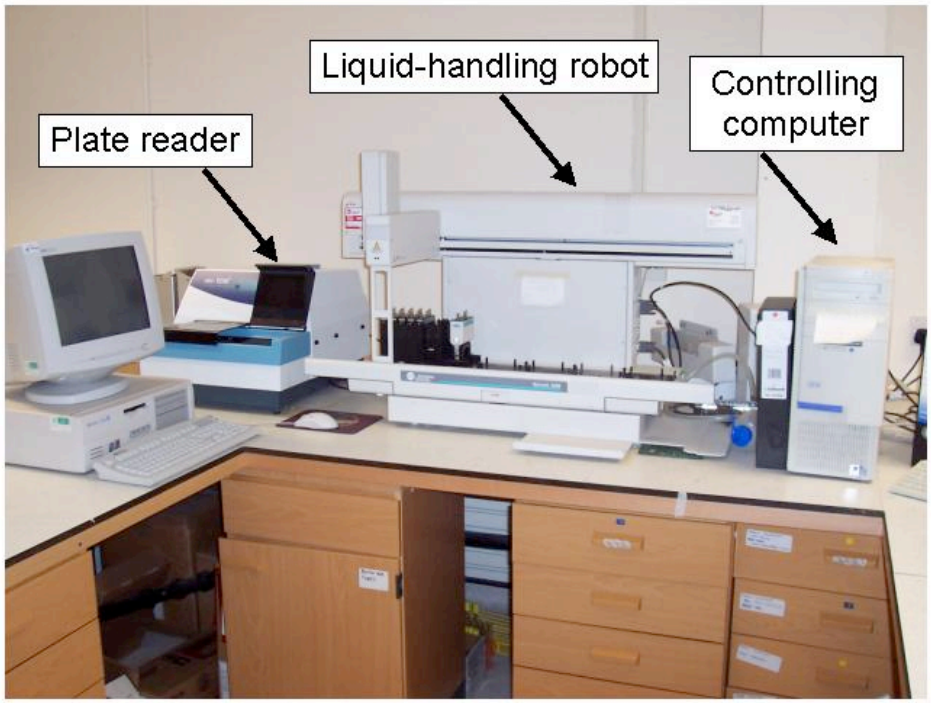
Auxotrophic Growth Experiments



experiment	growth?
ΔYPR060C (knockout)	no
ΔYPR060C + chorismate	no
ΔYPR060C + prephenate	yes
ΔYPR060C + phenylpyruvate	no

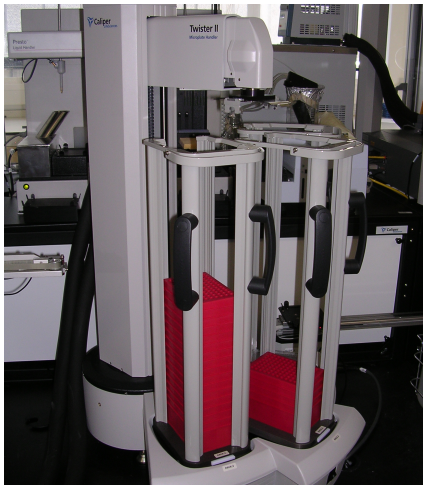


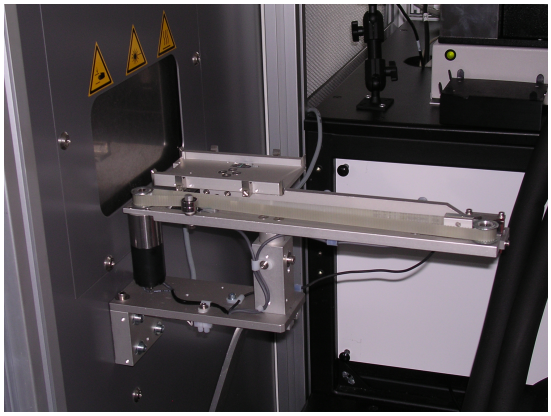
The Original Robot Scientist System



The New Robot Scientist Lab

The New Robot Scientist Lab



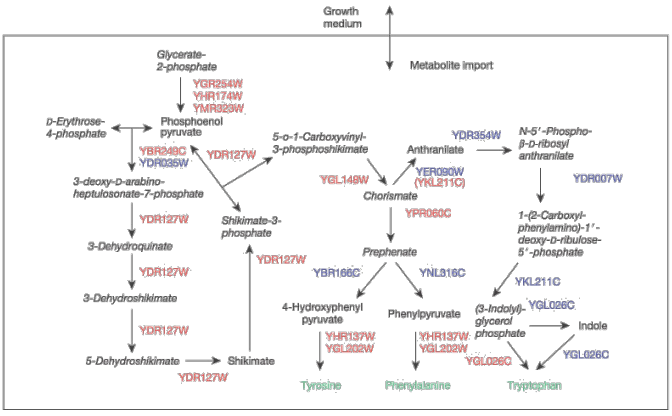


The Robot Scientist in Action

- videos available at

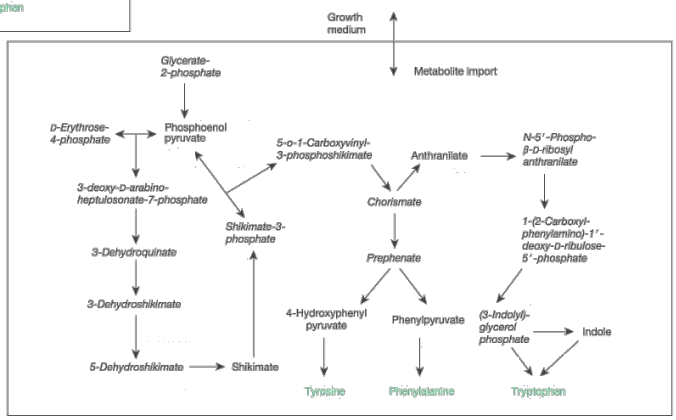
  http://www.aber.ac.uk/en/cs/research/cb/projects/robotscientist/video/

The Network Reconstruction Task



⇐ pathways for synthesizing aromatic amino acids: nodes in the graph are metabolites, edges are enzymes

King et al. assume that we have pathway model but do not know which genes encode which enzymes – goal is to infer this mapping

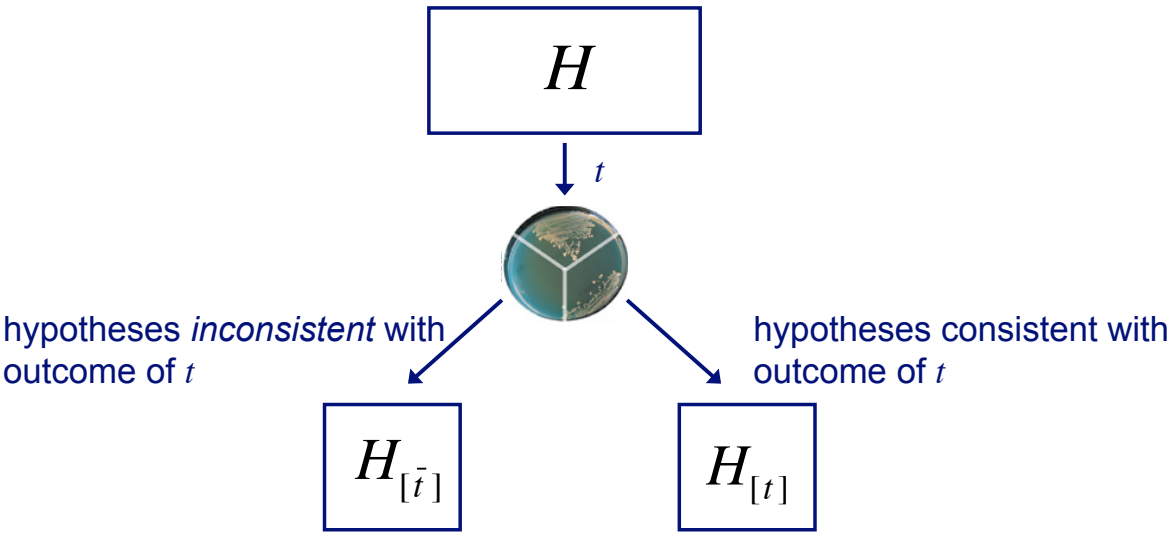


Selecting an Experiment in ASE-Progol

- on each *trial* the robot scientist can select
  - a knockout strain (i.e. a yeast strain with one particular gene disabled)
  - a growth medium
- different experiments have different costs (the costs of reagents varies by orders of magnitude)
- how should the system select the next experiment to run?
- goal is to find the correct model while minimizing the cost to do so

Selecting an Experiment

- given a set of candidate hypotheses H, and a trial t, the outcome of partitions the hypotheses into two sets

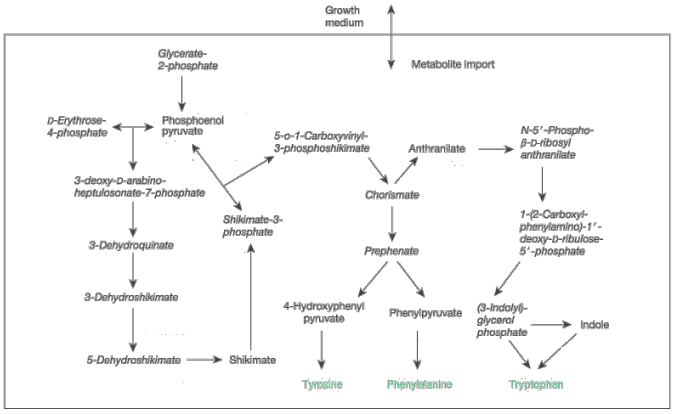


Hypotheses

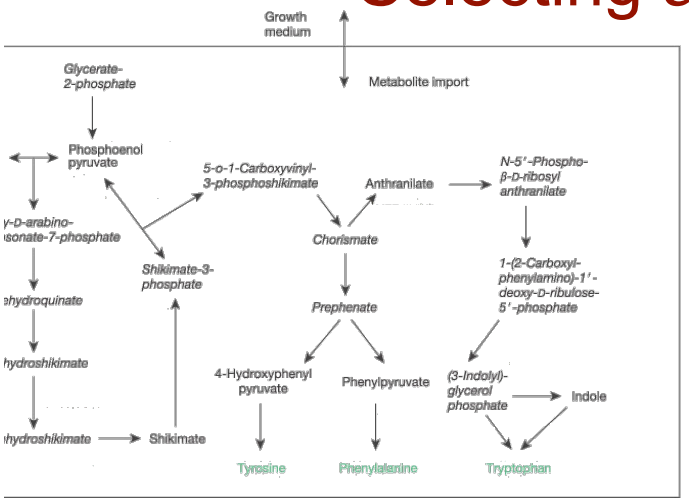
- hypotheses consist of assignments of genes to the reactions they catalyze

Prephenate
↓ YPR060C?
4-H pyruvate

Chorismate
↓ YPR060C?
Prephenate

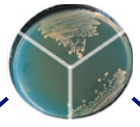


Selecting an Experiment



H

↓ t = ΔYPR060C + Prephenate



growth!

$H_{[\bar{t}]}$ Prephenate
↓ YPR060C?
4-H pyruvate

$H_{[t]}$ Chorismate
↓ YPR060C?
Prephenate

Selecting an Experiment

- given a set of candidate hypotheses H, and a set of candidate trials T, the minimum *expected cost* of experimentation is:

$$EC(H,T) = \min_{t \in T} \left[\begin{array}{l} C_t + p(t)\ EC(H_{[t]}, T - t) + \\ \qquad (1 - p(t))\ EC(H_{[\bar{t}]}, T - t) \end{array} \right]$$

- where

  C_t is the cost of trial t

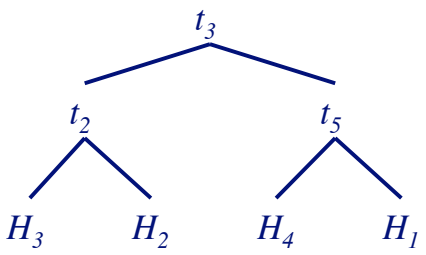
  $p(t)$ is the probability that the outcome of trial t is positive

  $EC(\varnothing, T) = 0$

  $EC(\{h\}, T) = 0$

Selecting an Experiment

- figuring out the optimal sequence of trials is equivalent to finding a minimum-cost decision tree which is NP-hard



- so we need an approximation…

Selecting an Experiment

- recall that in an optimal coding scheme, the number of bits to use for message h that has probability $p(h)$ is:

$$-\log_2(p(h))$$

- interpreting the bits of the code as outcomes of binary trials, the number of trials to eliminate all hypotheses except h is at most:

$$\lceil -\log_2(p(h)) \rceil$$

Selecting an Experiment

- given a set of hypotheses, and an estimated probability of each being true, we can calculate the expected number of trials to identify the correct hypothesis

$$J_H = -\sum_{h \in H} p(h) \log_2(p(h))$$

- this is the entropy of the distribution over hypothesis probabilities

Selecting an Experiment

- King et al. use the following approximation

$$EC(H,T) \approx \min_{t \in T} \left[\begin{array}{l} C_t + p(t)\ (\text{mean}_{t' \in (T-t)} C_{t'})\ J_{H_{[t]}} + \\ \qquad (1 - p(t))(\text{mean}_{t' \in (T-t)} C_{t'})\ J_{H_{[\bar{t}]}} \end{array} \right]$$

where

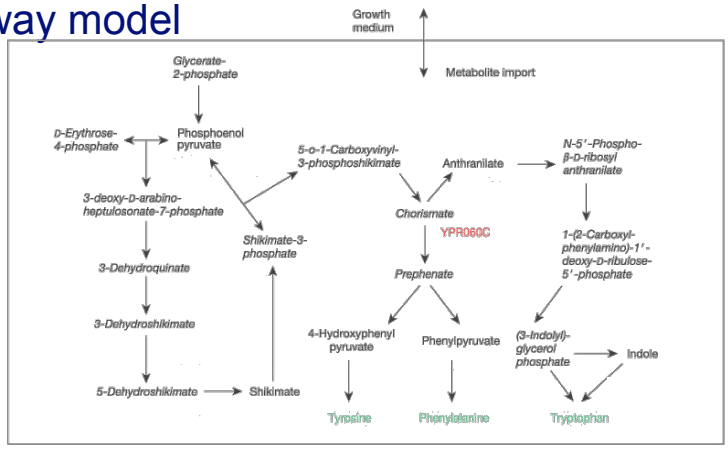
$$J_H = -\sum_{h \in H} p(h) \lfloor \log_2(p(h)) \rfloor$$

and $p(h)$ is the probability that hypothesis h is correct

The Logical Model in ASE-Progol

- get $p(h)$ by using logical inference to determine how well h "compresses" (explains) the observations so far
- get $p(t)$ by summing $p(h)$ for h that predict a positive result
- determine prediction for each hypothesis by doing logical inference on the pathway model

Will ΔYPR060C + prephenate grow given our hypothesis about YPR060C?

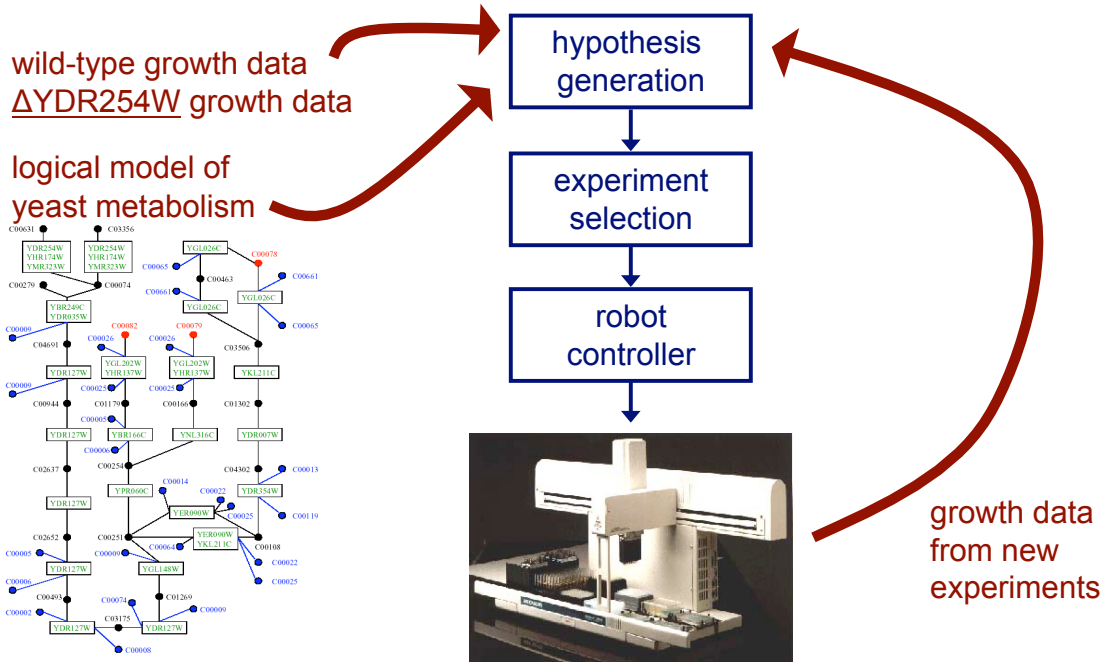


The ASE-Progol Main Loop

while $|H| > 1$
  for each candidate experiment t // may only be a sample
    for each candidate hypothesis h in H // may only be a sample
      determine h's prediction for t
    determine expected experimentation cost if we run t
  run trial that leads to min estimated experimentation cost

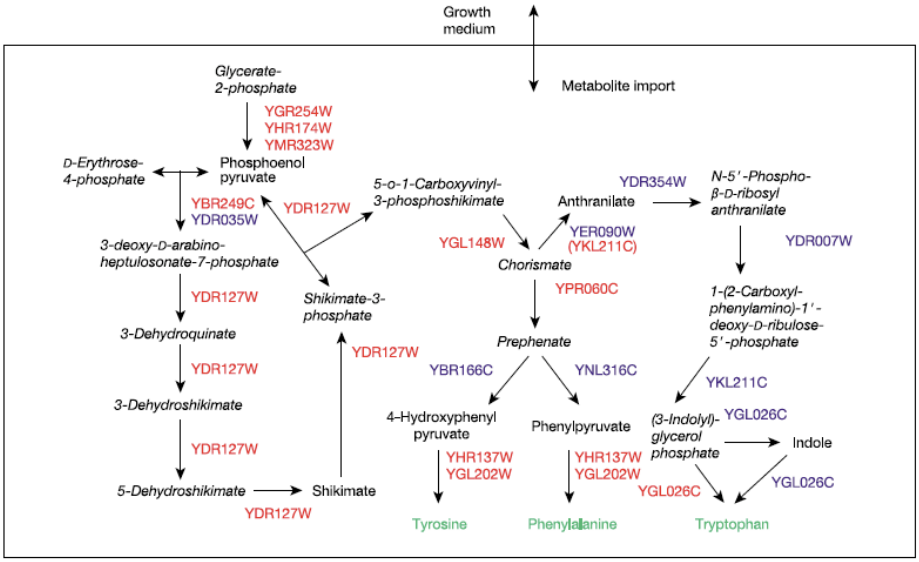
Active Experimentation with the Robot Scientist

What does the enzyme encoded by the ORF YDR254W do?



Experimental Evaluation (*Nature* 2004)

- try to reconstruct pathway model for synthesis of aromatic amino acids
- determine which genes are associated with which enzymatic reactions in the model



Experimental Evaluation (*Nature* 2004)

- compare three trial selection strategies
  - ASE
  - naïve: select the cheapest experiment not yet done
  - random

- evaluate accuracy of an approach by
  - considering predictions made for all single-metabolite and double-metabolite experiments
  - averaged over all hypotheses not eliminated

Experimental Evaluation (*Nature* 2004)

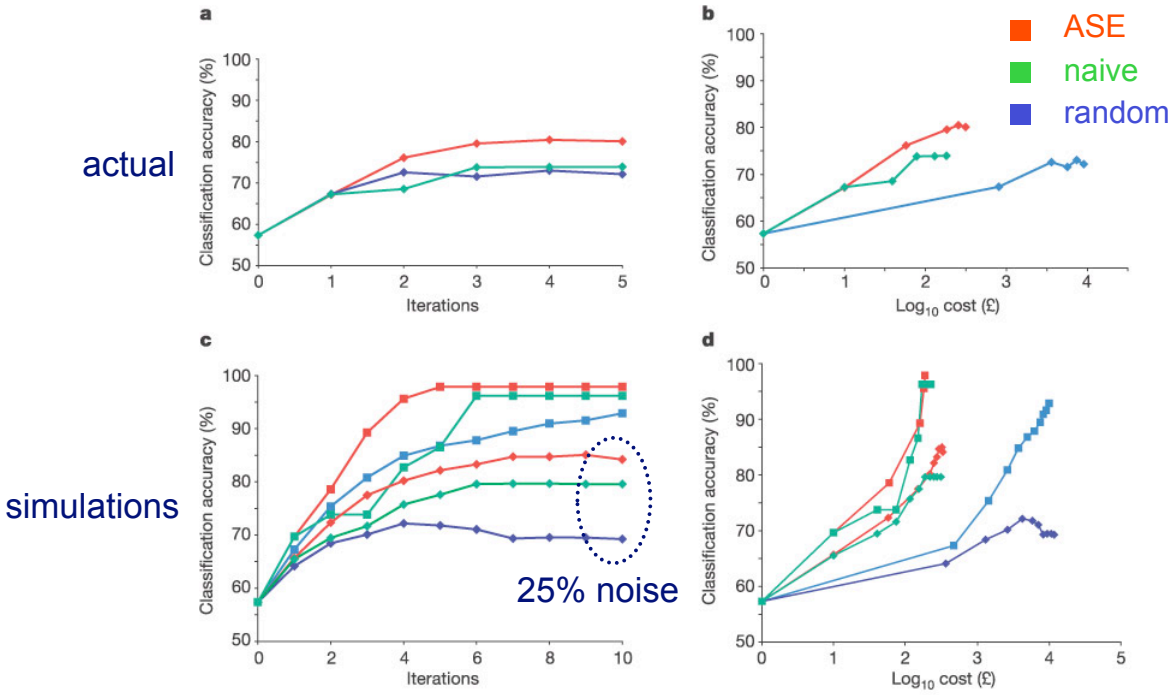
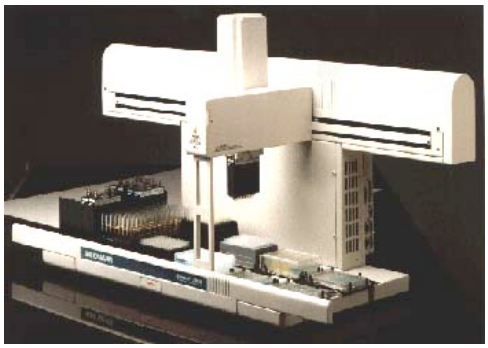


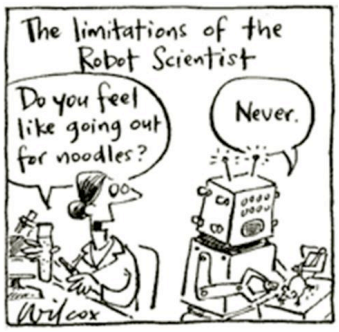
Figure from King et al., *Nature* 427:247-252, 2004.

Experimental Evaluation (*Nature* 2004)



vs.

"In initial trials, using nine graduate computer scientists and biologists, we found that there was no significant difference between the robot and the *best* human performance in terms of the number of iterations required to achieve a given level of accuracy."



Experimental Evaluation (*Science* 2009)

- identify genes encoding *orphan* enzymes (we know reaction occurs, we don't know which gene carries it out)
- scale of experiment
  - logical model of metabolism encodes ~1200 genes, ~800 metabolites
  - system made 6,657,024 optical-density measurements (each quantifies growth at a particular time point in a particular culture)

- investigated 20 hypotheses about 13 orphan enzymes
  - 12 hypotheses with no previous evidence were established with $p < 0.05$
  - confirmed with direct experimental methods

Experimental Evaluation (*Science* 2009)

Orphan enzyme	Hypothesized gene	Prob.	Acc.	No.	Existing annotation	Dry	Wet
Glucosamine-6-phosphate deaminase (3.5.99.6)	YHR163W (SOL3)	$<10^{-4}$	97	8	6-Phosphogluconolactonase, ida	–	–
Glutaminase (3.5.1.2)	YIL033C (BCY1)	$<10^{-4}$	92	11	Cyclic adenosine 3′,5′-monophosphate (cAMP)–dependent protein kinase inhibitor, ida	✗	–
L-Threonine 3-dehydrogenase (1.1.1.103)	YDL168W (SFA1)	$<10^{-4}$	83	6	Alcohol dehydrogenase, ida	–	–
Purine-nucleoside phosphorylase (2.4.2.1)	YLR209C (PNP1)	$<10^{-4}$	82	11	Purine-nucleoside phosphorylase, ida	✓	–
2-Aminoadipate transaminase (2.6.1.39)	YGL202W (ARO8)	$<10^{-4}$	80	3	Aromatic–amino acid transaminase, ida	✓	✓
5,10-Methenyltetrahydrofolate synthetase (6.3.3.2)	YER183C (FAU1)	$<10^{-4}$	80	4	5,10 Formyltetrahydrofolate cyclo-ligase, ida	✓	–
Glucosamine-6-phosphate deaminase (3.5.99.6)	YNR034W (SOL1)	$<10^{-4}$	79	2	Possible role in tRNA export	–	–
Pyridoxal kinase (2.7.1.35)	YPR121W (THI22)	$<10^{-4}$	78	1	Phosphomethylpyrimidine kinase, iss	–	–
Mannitol-1-phosphate 5-dehydrogenase (1.1.1.17)	YNR073C	$<10^{-4}$	78	6	Putative mannitol dehydrogenase, iss	–	–
1-Acylglycerol-3-phosphate O-acyltransferase (2.3.1.51)	YDL052C (SLC1)	0.0001	80	6	1-Acylglycerol-3-phosphate O-acyltransferase ida	✓	–