The Robot Scientist Project

Ross D. King, Michael Young, Amanda J. Clare, Kenneth E. Whelan,
and Jem Rowland

The University of Wales, Aberystwyth
{rdk, miy, afc, knw, jjr}@aber

Abstract. We are interested in the automation of science for both philosophical and technological reasons. To this end we have built the first automated system that is capable of automatically: originating hypotheses to explain data, devising experiments to test these hypotheses, physically running these experiments using a laboratory robot, interpreting the results, and then repeat the cycle. We call such automated systems "Robot Scientists". We applied our first Robot Scientist to predicting the function of genes in a well-understood part of the metabolism of the yeast *S. cerevisiae*. For background knowledge, we built a logical model of metabolism in Prolog. The experiments consisted of growing mutant yeast strains with known genes knocked out on specified growth media. The results of these experiments allowed the Robot Scientist to test hypotheses it had abductively inferred from the logical model. In empirical tests, the Robot Scientist experiment selection methodology outperformed both randomly selecting experiments, and a greedy strategy of always choosing the experiment of lowest cost; it was also as good as the best humans tested at the task. To extend this proof of principle result to the discovery of novel knowledge we require new hardware that is fully automated, a model of all of the known metabolism of yeast, and an efficient way of inferring probable hypotheses. We have made progress in all of these areas, and we are currently 6building a new Robot Scientist that we hope will be able to automatically discover new biological knowledge.

1 Introduction

1.1 The Robot Scientist Concept

The Robot Scientist project aims to develop computer systems that are capable of automatically: originating hypotheses to explain data, devising experiments to test these hypotheses, physically running these experiments using a laboratory robot, interpreting the results, and then repeat the cycle (Figure 1).

1.2 Motivation

- Philosophical - Our primary motivation is a better understanding of science. For us, the question of whether it is possible to automate the scientific discovery process is central to an understanding science, as we believe that we do not fully understand a phenomenon unless we can make a machine, which reproduces it.

A. Hoffmann, H. Motoda, and T. Scheffer (Eds.): DS 2005, LNAI 3735, pp. 16 – 25, 2005.

- Technical - In many areas of science our ability to generate data is outstripping our ability to analyse the data. One scientific area where this is true is post-genomic Biology where data is now being generated on an industrial scale. We contend that the analysis of scientific data needs to become as industrialized as its generation.

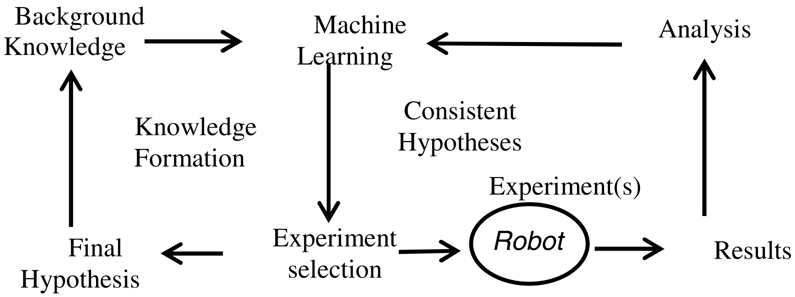

Fig. 1. The Robot Scientist Hypothesis Generation, Experimentation, and Knowledge Formation loops

1.3 Scientific Discovery

The branch of Artificial Intelligence devoted to developing algorithms for acquiring scientific knowledge is known as "scientific discovery". The pioneering work in the field was the development of learning algorithms for analysis of mass-spectrometric data [1]. This work was notable as an early example of interdisciplinary research: it involved world-class scientists from biology (J. Lederberg), chemistry (C. Djerassi), and computer science (E. Feigenbaum). This project initiated the whole field of machine learning. In the subsequent 30 years, much has been achieved, and there are now a number of convincing examples where computer programs have made explicit contributions to scientific knowledge [2,3]. However, the general impact of such programs on science has been limited. This is now changing, as the confluence of the expansion of automation in science, and advances in AI, are making it increasingly possible to couple scientific discovery software with laboratory instrumentation.

2 Previous Work on the Robot Scientist

In [4] we first developed the Robot Scientist concept. The Robot Scientist is a reasoned, but radically new, approach to scientific discovery that seeks to integrate data generation and analysis in a physically closed loop. A widely accepted view of science is that it follows a "hypothetico-deductive" process [5]. Scientific expertise and imagination are first used to form possible hypotheses, and then the deductive consequences of these hypotheses are tested by experiment. The Robot Scientist methodology (Figure 1) is consistent with this paradigm: we employ the logical inference mechanism of abduction [6] to form new hypotheses, and that of deduction to test which hypotheses are consistent.

2.1 The Biological System

The first aim of the first Robot Scientist project was to develop a proof-of-principle system that would demonstrate automated cycles of hypothesis generation and experiment on a real biological system. For this we chose the scientific area of "Functional Genomics". The aim of this branch of biology is to both uncover the function of genes identified from sequencing projects (such as that on the human genome), and to better characterize the function of genes with currently putative functions. We chose to focus on brewer's (or baker's) yeast (*S. cerivisae*). This is the best understood eukaryotic organism. As humans are eukaryotic organisms, this yeast is used as a "model" of human cells, as it is simpler and easier to work with. *S. cerivisae* was the first eukaryotic organism sequenced and has been studied for over a hundred and fifty years. Despite this, around 30% of its ~6,000 genes still have no known function.

A key advantage with working with yeast is that it is possible to obtain strains of yeast with each of the ~6,000 genes knocked out (removed). We chose to use these mutatnts along with a classical genetic technique known as "auxotrophic growth experiments". These experiments consist of making particular growth media and testing if the mutants can grow (add metabolites to a basic defined medium). A mutant is auxotrophic if cannot grow on a defined medium that the wild type can grow on. By observing the pattern of metabolites that recover growth, the function of the knocked out mutant can be inferred. We focused on the aromatic amino acid (AAA) pathway in yeast.

2.2 Logical Model

In any scientific discovery problem that is not purely phenomenological we need to develop a model of the natural system. We therefore developed a logical formalism for modelling cellular metabolism that captures the key relationships between protein-coding sequences (genes, ORFs), enzymes, and metabolites in a pathway, along with feedback, etc. [6]. This model is expressed in predicate logic and encoded in the logic programming language Prolog (see Figure 2.).

Logic is our oldest (>2,500 years) and best understood way of expressing knowledge, and computer programs are the most general way we have of expressing knowledge: logic programs combine the clarity of logic with the expressive power of computer programs. All objects (genes, proteins, metabolites) and their relationships (coding, reactions, transport, feed-back) are described as logical formulae. The structure of the metabolic models pathway is that of directed graphs, with metabolites as nodes and enzymes as arcs. An edge arc corresponds to a reaction. The compounds at each vertex node are the set of all metabolites and the compounds that can be synthesised by the reactions leading to it. Reactions are modelled as unidirectional transformations. A model's consistency and completeness can be analysed by comparing the model's logical consequences with the outcomes of *in vivo* auxotrophic growth experiments. The model can thus be used to yield a procedural specification of the functional genomics problem, namely how to infer gene functions from experimental observations. The model is both declarative (expressing text-book biochemistry) and procedural (enabling inferences about pathways). In particular, two types of inference can be made: deductions to infer phenotype, and abductions to infer missing reactions (gene functions) see Figure 3. A mutant is inferred (deduced) to grow if and only if, a path can be found from the input metabolites to the three aromatic amino acids. Conversely, a mutant is inferred to be auxotrophic if, and only if, no such path can be found. We formed our "gold-standard" AAA model to fit both the existing knowledge on the AAA pathway and our experimental auxotrophic growth experiments.

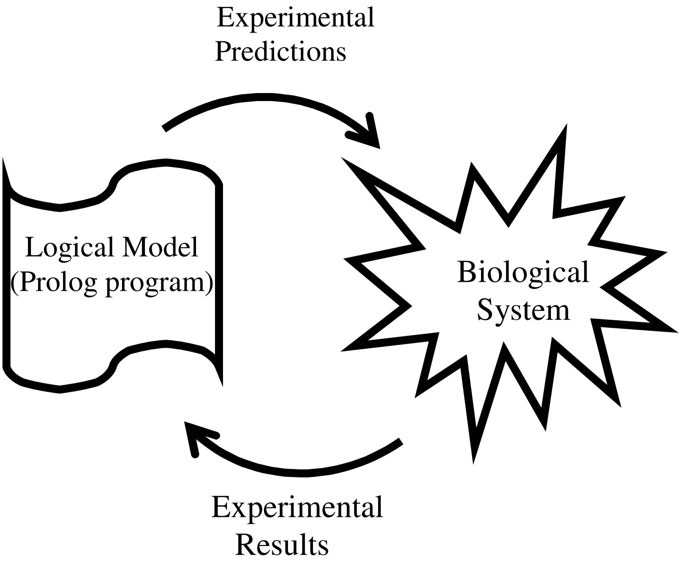

Fig. 2. The Relationship between the logical model and the experimental System

Deduction

Rule: If a cell grows, then it can synthesise tryptophan.
Fact: cell cannot synthesise tryptophan
$\therefore$ Cell cannot grow.

Given the rule $P \rightarrow Q$, and the fact $\neg Q$, infer the fact $\neg P$ (*modus tollens*)

Abduction

Rule: If a cell grows, then it can synthesise tryptophan.
Fact: Cell cannot grow.
$\therefore$ Cell cannot synthesise tryptophan.

Given the rule $P \rightarrow Q$, and the fact $\neg P$, infer the fact $\neg Q$

Fig. 3. Simplified form of the deductive and abductive inference used

The form of the hypotheses that were abductively inferred was very simple. Each hypothesis binds a particular gene to an enzyme that catalyses the reaction. For example:

- A correct hypothesis would be that: YDR060C codes for the enzyme for the reaction: chorismate → prephenate.
- An incorrect hypothesis would be that: it coded for the reaction: chorismate → anthranilate.

2.3 Active Learning

The branch of machine learning that deals with algorithms that can choose their own examples (experiments) is known as "active learning" [7]. If we assume that each hypothesis has a prior probability of being correct, and that each experiment has an associated price, then scientific experiment selection can be formalised as the task of: given a set of possible hypotheses, each with a probability of being correct, and given that each experiment has an associated cost, select the optimal series of experiments (in terms of expected cost) to eliminate all but the one correct hypothesis [8]. This problem is, in general, computationally intractable (NP-hard). However, it can be shown that the experiment selection problem is structurally identical to finding the smallest decision tree, where experiments are nodes, and hypotheses leaves. This is significant because a Bayesian analysis of decision-tree learning has shown that near-optimal solutions can be found in polynomial time [9]. To approximate the full Bayesian solution we use the following [8]. *EC(H,T)* denote the minimum expected cost of experimentation given the set of candidate hypotheses *H* and the set of candidate trials *T*:

$$EC(\varnothing, T) = 0$$

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

$$EC(H,T) \approx \min_{t \in T}\left[C_t + p(t)(mean_{t' \in (T-t)} C_{t'})J_{H[t]} + (1 - p(t))mean_{t' \in (T-t)} C_{t'})J_{H[\bar{t}]}\right]$$

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

Ct is the monetary price of the trial *t*

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

p(t) can be computed as the sum of the probabilities of the hypotheses (*h*) which are consistent with a positive outcome of *t*.

2.4 Results

A Robot Scientist was physically implemented that can conduct biological assays with minimal human intervention after the robot is set up [4]. The hardware platform consisted of a liquid-handling robot (Biomek 2000) with its control PC, a plate reader (Wallac 1420 Multilabel counter) with its control PC, and a master PC to control the system and do the scientific reasoning. The software platform consisted of background knowledge about the biological problem, a logical inference engine, hypothesis generation code (abduction), experiment selection code (deduction), and the Laboratory Information Management System (LIMS) code that glued the whole system together. We used the Inductive Logic Programming (ILP system ASE-Progol. The robot conducted experiments by pipetting and mixing liquids on microtitre plates. Given a computed definition of one or more experiments, we developed code which designed a layout of reagents on the liquid-handling platform that would enable these experiments, with controls, to be carried out efficiently. In addition, the liquid-handling robot was automatically programmed to plate out the yeast and media into the correctly specified wells. The system measured the concentration of yeast in the wells of the microtitre trays using the adjacent plate reader and returns the results to the LIMS (although microtitre trays were still moved in and out of incubators manually). *The key point is that there was no human intellectual input in the design of experiments or the interpretation of data.*

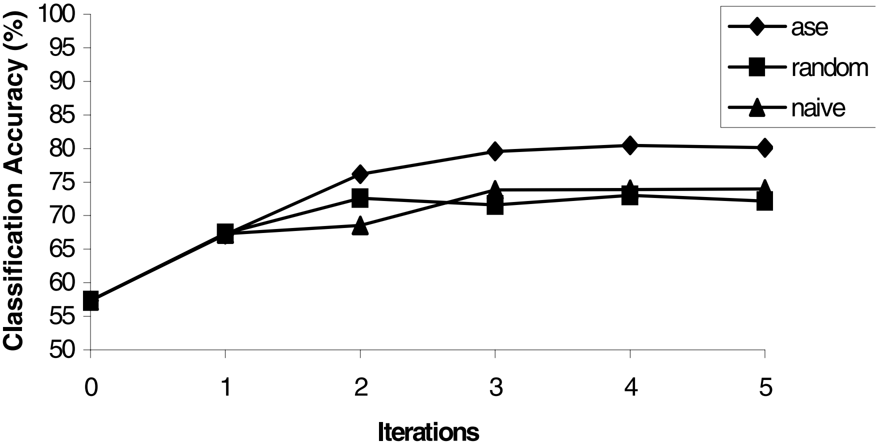

Fig. 4. The observed classification accuracy versus iterations (time)

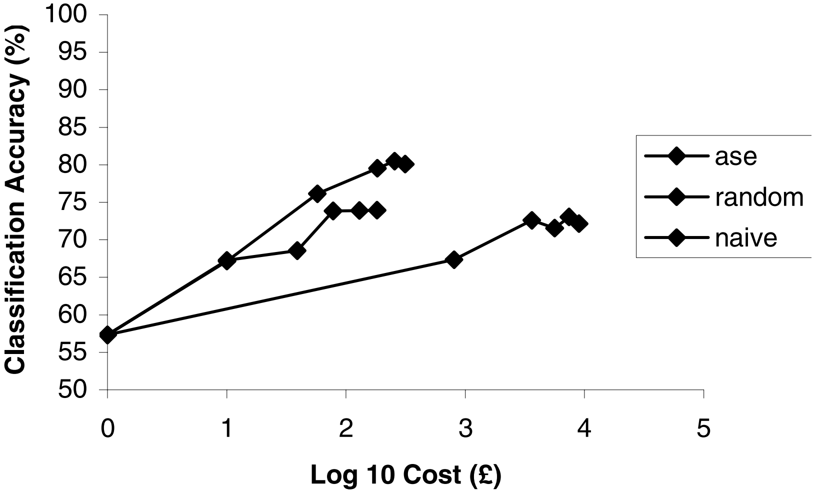

Fig. 5. The Observed classification accuracy versus cost (price of chemicals)

Figures 4 (below) shows the average classification accuracy versus experimental iteration (time) for the robot's intelligent strategy (red) compared with random (blue), and the naïve strategy of always choose the cheapest experiment (green). Figure 5 shows the average classification accuracy versus money spent (£). The intelligent strategy is both significantly faster and cheaper than the other strategies. When compared with human performance on this task, the Robot was as good as the best human scientists.

3 Current Status of the Robot Scientist

In our first work with the Robot Scientist we demonsrtaed a proof-of-principle application of the Robot Scientist. We demonsrtaed that we could automatically *rediscover* known biological knowledge. We now wish to extend this result to the *discovery* of new biological knowledge. To achieve this we have chosen to focus on the same biological problem. However to actually fully automattedly discover new knowledge a number of extensions are required:

- New Hardware
  - The original hardware was not fully automated, and several steps had to be done manually at the request of the Robot Scientist. We wish to make the system fully automated.
  - The experimental throughput capacity of the original hardware was also limited. A key advantage of automation is that it can be scaled up. The new hardware will have far greater capacity.
  - We will also extend the original qualitative experimental methodology (growth v no-growth) to a quantitative measurement of growth.
- Expansion of the background knowledge to include as much as possible of what is known about yeast metabolism. For if the Robot Scientist does not already know what is scientifically known, then it will be very difficult to discover something novel. This will require a move from a model with ~10 reactions to a model with more than 1,000 reactions. Our current model includes 1,166 genes (940 known, 226 inferred). As in the original AAA model, growth is predicted if there exists a path from the growth medium to defined end-points.
- Improve the efficiency of the hypothesis generation method. The current approach is purely logical and does not take advantage of domain knowledge. This approach will not scale to a model of two orders of magnitude greater size. Therefore, we will use bioinformatics to incorporate biological knowledge. One way of thinking about current bioinformatic genome annotation is as *hypothesis formation processes*; and hypothesis formation is perhaps the hardest part of automating science. Therefore, bioinformatic methods will generate the hypotheses that the robot scientist will experimentally test.

3.1 The New Robot Scientist Hardware

Our new Robotic Scientist hardware will be commissioned in the last quarter of 2005, and will cost £450,000 (see Figure 6). It will be manufactured by Caliper Life

Sciences. The hardware will consist of the following components: -80C freezer, a liquid-handling robot, incubator(s), plate-reader(s), and robot arms. The robotic system is designed to be able to be able to in a completely automatic manner: select frozen yeast strains from the freezer, inoculate these strains into a rich medium, harvest a defined quantity of cells, inoculate these cells into specified media (base plus added metabolites and/or inhibitors), and accurately measure growth in the specified media. Design of this system has been extremely challenging, and the specification has taken over 6 months to refine and make practical. To the best of our knowledge, after extensive discussions with manufacturers, we are confident that there is no comparable system anywhere in the world that can flexibly automate anything close to as many growth experiments. The system will be capable of initiating >1,000 new strain/defined growth-medium experiments a day, and each experiment will last up to 3days (plus an initiation day), using a minimum of 50 different yeast strains. It will be possible to take an optical density (OD) measurement for an experiment every 20 minutes, enabling accurate growth curves to be formed. It will also be possible to take samples from experiments for more detailed analysis, or to inoculate other experiments. The system will be able to run "lights out" for days at a time.

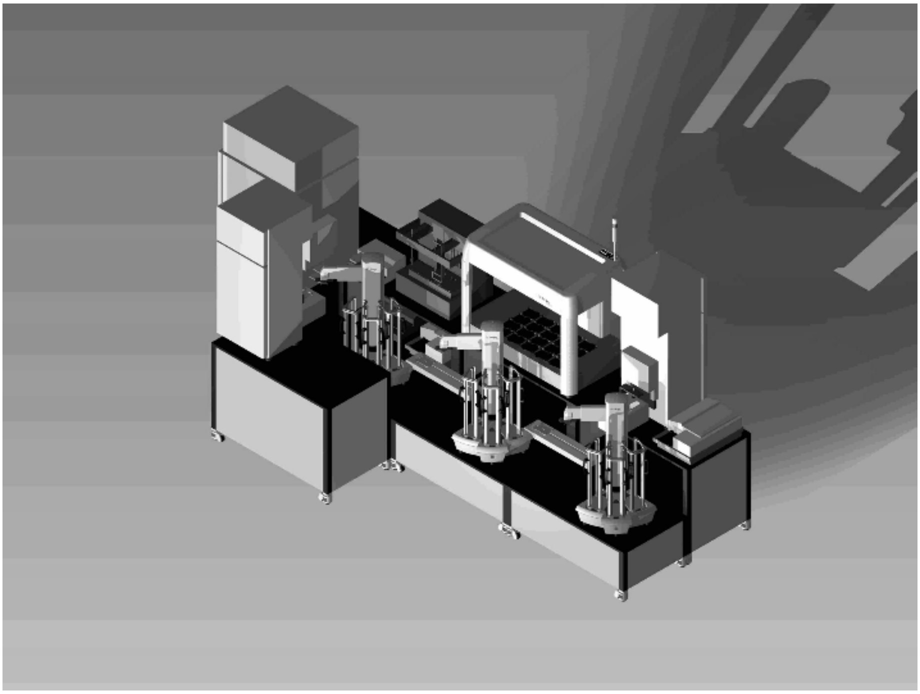

Fig. 6. Sketch of the new Robot Scientist hardware

The class of the experiments possible using this new hardware is comparable to those that the Robot Scientist can currently undertake. However the major advances will be:

- A huge increase in the scale of the number of experiments performed. Using our existing robotic system, we can perform ~200 strain/medium growth measurements a day: with our new robotic system we will be able to perform >100,000.
- A reduction in experimental noise. The current laboratory robot has ~25% noise when assaying growth or no-growth - mostly due to it being in a non-sterile environment, and cross-plate contamination. This noise will be drastically reduced, increasing throughput (through fewer controls being required), and simplifying data analysis.
- Accurate quantitative measurement of growth. Most genes display quantitative rather than qualitative effects under most environmental conditions [*Ol2*].
- Measure growth curves and yield.
- An increase in the range of metabolites used. We plan to have ~500 metabolites available, compared to a ~50 at present.
- The use of specific enzyme inhibitors.
- An increase in the range of strains used: including a set of Canadian double knockouts, and a set of knock-down mutants: where essential genes have been placed under the control of a promoter (e.g. *tet*O).
- The experiments will be fully automatic. Currently the Robot Scientist needs to direct a technician to execute a number of steps.

3.2 The Experimental Plan

The plan is to initiate ~1,000 experiments a day, providing ~200,000 daily measurements (based on a 3-day cycle measuring every 20 mins.). The reason that so many experiments are required is that even relatively simple cells, such as those of *S. cerevisiae*, are extremely complicated systems involving thousands of genes, proteins, and metabolites. Such systems can be in astronomical numbers of states, and the only possible way to dissect them is to be intelligent, and to do large numbers of experiments. One way to think about it is as an information theory problem: a complicated message cannot be sent using a few bits. Note, we do not plan to do all possible experiments, as even to test all possible pairs of metabolites would involve: 6,000 (genes) * (500 (metabolites))^2 = 1,500,000,000 (experiments).

All results will be stored in the Bio-Logical relational database (see above) along with meta-data detailing the experimental conditions. We expect to produce >40,000,000 growth measurements and all these results will be placed in the public domain. On their own, these results will constitute a significant contribution to scientific knowledge. N.B. the existing bioinformatic information on the growth of knockouts is often very poor, i.e. often gene knock-outs labelled as “lethal” have no description of the growth medium used, and the information is often also unreliable as it was produced using noisy high-throughput screens.

4 Discussion

The general Robot Scientist idea could be applied to many scientific problems. We are actively investigating the following two areas:

- Drug design - selection of compounds from libraries and/or use of laboratory on chip technology. The idea here is to incorporate the Robot Scientist into a Quantitative Structure Activity Relationship (QSAR) system [10].
- Quantum control - using femtosecond (10^{-15}s) lasers to control chemical synthesis. We are collaborating with the Department of Chemistry at the University of Leeds (UK) to use Robot Scientist type ideas to control the search for patterns of femtosecond laser pulses that can act as "optical catalysts" [11]. The main difference with this application and those in yeast is that the experiments take ~1s (and could be as low as 0.001s), compared to 24hours with yeast.

Acknowledgements

We are grateful for support from the BBSRC (2/E11552, E20268) and HEFCW (Aber 4). Mike Benway of Claiper Life Sciences made an important contribution to the new the system design.

References

1. Buchanan, B.G., Sutherland, G.L., Feigenbaum, E.A. Toward automated discovery in the biological sciences. In Machine Intelligence 4, Edinburgh University Press, (1969) 209-254
2. Langley, P., Simon, H.A., Bradshaw, G.L., Zytkow, J.M.: Scientific Discovery: Computational Explorations of the Creative Process. MIT Press (1987)
3. King, R.D., Muggleton, S.H., Srinivasan, A., Sternberg, M.J.E.: Structure-activity relationships derived by machine learning: The use of atoms and their bond connectivities to predict mutagenicity by inductive logic programming. Proc. Nat. Acad. Sci. USA 93, (1996) 438-442
4. King R.D, Whelan K.E, Jones F.M, Reiser P,J,K Bryant C.H, Muggleton S, Kell D.B, Oliver S.: Functional Genomic Hypothesis Generation and Experimentation by a Robot Scientist. Nature 427 (1994) 247-252
5. Popper, K. The Logic of Scientific Discovery, Hutchinson, London (1972)
6. Reiser, P.G.K., King, R.D., Kell, D.B., Muggleton, S.H., Bryant, C.H. Oliver, S.G.: Developing a logical model of yeast metabolism. Electronic Transactions in Artificial Intelligence 5, (2001) 223-244
7. Duda, R.O., Hart, P.E., Stork, D.G. Pattern Classification. Wiley. (2001)
8. Bryant, C.H., Muggleton, S.H., Oliver, S.G., Kell, D.B., Reiser, P, King, R.D.: Combining inductive logic programming, active learning, and robotics to discover the function of genes. Electronic Transactions in Artificial Intelligence. 5 (2001) 1-36
9. Muggleton, S. Page, D.: A learnability model of universal representations and its application to top-down induction of decision trees. In K. Furukawa, D. Michie, & S. Muggleton (eds.) Machine Intelligence 15, Oxford University Press, (1999) 248-267.
10. King, R.D., Muggleton, S., Lewis R.A., Sternberg, M.J.E.: Drug design by machine learning: The use of inductive logic programming to model the structure-activity relationships of trimethoprim analogues binding to dihydrofolate reductase. Proc. Nat. Acad. Sci. U.S.A. 89. (1992) 11322-11326.
11. Levis, R.J., Menkir, G., Rabitz H.: Selective Covalent Bond Dissociation and Rearrangement by Closed-Loop, Optimal Control of Tailored, Strong Field Laser Pulses, Science, 292, (2001)709-713