# Computational Discovery of Scientific Knowledge

Sašo Džeroski[1], Pat Langley[2], and Ljupčo Todorovski[1]

[1] Department of Knowledge Technologies, Jozef Stefan Institute
Jamova 39, SI-1000 Ljubljana, Slovenia
Saso.Dzeroski@ijs.si, Ljupco.Todorovski@ijs.si
[2] Computational Learning Laboratory
Center for the Study of Language and Information
Stanford University, Stanford, CA 94305 USA
langley@isle.org

**Abstract.** This chapter introduces the field of computational scientific discovery and provides a brief overview thereof. We first try to be more specific about what scientific discovery is and also place it in the broader context of the scientific enterprise. We discuss the components of scientific behavior, that is, the knowledge structures that arise in science and the processes that manipulate them. We give a brief historical review of research in computational scientific discovery and discuss the lessons learned, especially in relation to work in data mining that has recently received substantial attention. Finally, we discuss the contents of the book and how it fits in the overall framework of computational scientific discovery.

## 1 Introduction

This book deals with computational approaches to scientific discovery. Research on computational scientific discovery aims to develop computer systems which produce results that, if a human scientist did the same, we would refer to as discoveries. Of course, if we hope to develop computational methods for scientific discovery, we must be more specific about the nature of such discoveries and how they relate to the broader context of the scientific enterprise.

The term science refers both to scientific knowledge and the process of acquiring such knowledge. It includes any systematic field of study that relates to observed phenomena (as opposed to mathematics) and that involves claims which can be tested empirically (as opposed to philosophy). We will attempt to characterize science more fully later in the chapter, but one thing is clear: Science is about knowledge.

Science is perhaps the most complex human intellectual activity, which makes it difficult to describe. Shrager and Langley (1990) analyze it in terms of the knowledge structures that scientists consider and the processes or activities they use to transform them. Basic knowledge structures that arise in science include observations, laws, and theories, and related activities include data collection, law formation, and theory construction.

S. Džeroski and L. Todorovski (Eds.): Computational Discovery, LNAI 4660, pp. 1–14, 2007.

There are two primary reasons why we might want to study scientific discovery from a computational perspective:

- to understand how humans perform this intriguing activity, which belongs to the realm of cognitive science; and
- to automate or assist in facets of the scientific process, which belongs to the realm of artificial intelligence.

Science is a highly complex intellectual endeavor, and discovery is arguably the most creative part of the scientific process. Thus, efforts to automate it completely would rightfully be judged as audacious, but, as Simon (1966) noted, one can view many kinds of scientific discovery as examples of problem solving through heuristic search. Most research in automating scientific discovery has focused on small, well-defined tasks that are amenable to such treatment and that allow measurable progress.

Traditional accounts of science (Klemke et al., 1998) focus on the individual, who supposedly observes nature, hypothesizes laws or theories, and tests them against new observations. Most computational models of scientific discovery share this concern with individual behavior. However, science is almost always a collective activity that is conducted by interacting members of a scientific community. The most fundamental demonstration of this fact is the emphasize placed on communicating one's findings to other researchers in journal articles and conference presentations.

This emphasis on exchanging results makes it essential that scientific knowledge be *communicable*. We will not attempt to define this term, but it seems clear that contributions are more communicable if they are cast in established formalisms and if they make contact with concepts that are familiar to most researchers in the respective field of study. The research reported in this book focuses on computational discovery of such communicable knowledge.

In the remainder of this chapter, we first examine more closely the scientific method and its relation to scientific discovery. After this, we discuss the components of scientific behavior, that is, the knowledge structures that arise in science and the processes that manipulate them. We then give a brief historical review of research in computational scientific discovery and discuss the lessons learned, especially in relation to work in data mining that has recently received substantial attention. Finally, we discuss the contents of the book and how it fits in the overall framework of computational scientific discovery.

## 2 The Scientific Method and Scientific Discovery

The Merriam-Webster Dictionary (2003) defines science as: "a) knowledge or a system of knowledge covering general truths or the operation of general laws, especially as obtained and tested through the scientific method, and b) such knowledge or such a system of knowledge concerned with the physical world and its phenomena". The scientific method, in turn, is defined as the "principles and procedures for the systematic pursuit of knowledge involving the recognition

and formulation of a problem, the collection of data through observation and experiment, and the formulation and testing of hypotheses".

While there is consensus that science revolves around knowledge, there are different views in the philosophy of science (Klemke et al., 1998; Achinstein, 2004) about the nature of its content. The 'causal realism' position is that scientific knowledge is ontological, in that it identifies entities in the world, their causal powers, and the mechanisms through which they exert influence. In contrast, the 'constructive empiricism' tradition states that, scientific theories are objective, testable, and predictive. We believe that both frameworks are correct, in that they describe different facets of the truth.

The *scientific method* (Gower, 1996), dedicated to the systematic pursuit of reliable knowledge, incorporates a number of steps. First we must ask some meaningful question or identify a significant problem. We must next gather information relevant to the question, which might include existing scientific knowledge or new observations. We then formulate a hypothesis that could plausibly answer the question.

Next we must test this proposal by making observations and determining whether they are consistent with the hypothesis' predictions. When observations are consistent with the hypothesis, they lend it support and we may consider publishing it. If other scientists can reproduce our results, then the community comes to consider it as reliable knowledge. In contrast, if the observations are inconsistent, we should reject the hypothesis and either abandon it or, more typically, modify it, at which point the testing process continues. Hypotheses can take many different forms, including taxonomies, empirical laws, and explanatory theories, but all of them can be evaluated by comparing their implications to observed phenomena.

Most analyses of the scientific method come from philosophers of science, who have focused mainly on the evaluation of hypotheses and largely ignored their generation and revision. Unfortunately, what we refer to as discovery resides in just these activities. Thus, although there is a large literature on normative methods for making predictions from hypotheses, checking their consistency, and determining whether they are valid, there are remarkably few treatments of their production. Some (e.g., Popper (1959)) have even suggested that rational accounts of the discovery process are impossible. A few philosophers (e.g., Darden (2006); Hanson (1958); Lakatos (1976)) have gone against this trend and made important contributions to the topic, but most efforts have come from artificial intelligence and cognitive science.

Briefly, scientific discovery is the process by which a scientist creates or finds some hitherto unknown knowledge, such as a class of objects, an empirical law, or an explanatory theory. The knowledge in question may also be referred to as a scientific discovery. An important aspect of many knowledge structures, such as laws and theories, is their generality, in that they apply to many specific situations or many specific observations. We maintain that generality is an essential feature of a meaningful discovery, as will become apparent in the next section when we discuss types of scientific knowledge.

A defining aspect of discovery is that the knowledge should be new and previously unknown. Naturally, one might ask 'new to whom?'. We take the position that the knowledge should be unknown to the scientist in question with respect to the observations and background knowledge available to him when he made the discovery. This means that two or more scientists can make the same discovery independently, sometimes years apart, which has indeed happened in practice many times throughout the history of science. In this view, scientific discovery concerns a change in an individual's knowledge, which means that developing computer systems that reproduce events from the history of science can still provide important insights into the nature of discovery processes.

## 3 The Elements of Scientific Behavior

To describe scientific behavior, we follow Shrager and Langley (1990) and use as basic components knowledge structures and the activities that transform them. The former represent the raw materials and products of science, while the latter concern the process of producing scientific knowledge. The account below mostly follows the earlier treatise, but the definitions of several knowledge structures and activities have changed, reflecting improvements in our understanding over the past 15 years.

### 3.1 Scientific Knowledge Structures

Science is largely about understanding the world in which we live. To this end, we gather information about the world. Observation is the primary means of collecting this information, and observations are the primary input to the process of scientific discovery.

Observations (or data) represent recordings of the environment made by sensors or measuring instruments. Typically, the state of the environment varies over time or under different conditions, and one makes recordings for these different states, where what constitutes a state depends on the object of scientific study. We will refer to each of these recordings as an observation.

We can identify three important types of scientific knowledge – taxonomies, laws, and theories – that constitute the major products of the scientific enterprise. The creation of new taxonomies, laws, and theories, as well as revising and improving existing ones, make up the bulk of scientific discovery, making them some of the key activities in science.

- *Taxonomies* define or describe concepts for a domain, along with specialization relations among them. A prototypical example is the taxonomy for biological organisms, which are grouped into species, genera, families, and so forth, but similar structures play important roles in particle physics, chemistry, astronomy, and many other sciences. Taxonomies specify the concepts and terms used to state laws and theories.
- *Laws* summarize relations among observed variables, objects, or events. For example, Black's heat law states that mixing two substances produces a

temperature increases in one substance and a decrease in the other until they reach equilibrium. The law also describes a precise numeric relationship between the initial and final temperatures. The first statement is qualitative in form, whereas the latter is quantitative. Some laws may be quite general, whereas others may be very specific.

- *Theories* are statements about the structures or processes that arise in the environment. A theory is stated using terms from the domain's taxonomy and interconnects a set of laws into a unified theoretical account. For example, Boyle's law describes the inverse relation between the pressure and volume of a gas, whereas Charles' law states the direct relation between its temperature and pressure. The kinetic theory of gases provides a unifying account for both, explaining them in terms of Newtonian interactions among unobserved molecules.

Note that all three kinds of knowledge are important and present in the body of scientific knowledge. Different types of knowledge are generated at different stages in the development of a scientific discipline. Taxonomies are generated early in a field's history, providing the basic concepts for the discipline. After this, scientists formulate empirical laws based on their observations. Eventually, these laws give rise to theories that provide a deeper understanding of the structures and processes studied in the discipline.

A knowledge structure that a scientist has proposed, but that has not yet been tested with respect to observations, is termed an hypothesis. Note that taxonomies, laws, and theories can all have this status. As mentioned earlier, hypotheses must be evaluated to determine whether they are consistent with observations (and background knowledge). If it is consistent, we say that a hypothesis has been corroborated and it comes to be viewed as scientific knowledge. If an hypothesis is inconsistent with the evidence, then we either reject or modify it, giving rise to a new hypothesis that is further tested and evaluated.

Background knowledge is knowledge about the environment separate from that specifically under study. It typically includes previously generated scientific knowledge in the domain of study. Such knowledge differs from theories or laws at the hypothesis stage, in that the scientist regards it with relative certainty rather than as the subject of active evaluation. Scientific knowledge begins its life cycle as a hypothesis which (if corroborated) becomes background knowledge.

Besides the basic data and knowledge types considered above, several other types of structures play important roles in science. These include models, predictions, and explanations. These occupy an intermediate position, as they are derived from laws and theories and, as such, they are not primary products of the scientific process.

- *Models* are special cases of laws and theories that apply to particular situations in the environment and only hold under certain environmental conditions. These conditions specify the particular experimental or observational setting, with the model indicating how the law or theory applies in the setting. By applying laws and theories to a particular setting, models make it possible to use these for making predictions.

- *Predictions* represent expectations about the behavior of the environment under specific conditions. In science, a model is typically used to make a prediction, and then an actual observation is made of the behavior in the environment. Postdictions are analogous to predictions, except that the scientist generates them after making the observations he or she intends to explain. A prediction/postdiction that is consistent with the respective observation is successful and lends support to the model (and the respective law/theory) that produced it.
- *Explanations* are narratives that connect a theory to a law (or a model to a prediction) by a chain of inferences appropriate to the field. In such cases, we say that the theory explains the law. In some disciplines, inference chains must be deductive or mathematical. If a law cannot be explained by a theory (or a prediction by a model), we have an anomaly that brings either the theory or the observation into question.

### 3.2 Scientific Activities

Scientific processes and activities are concerned with generating and manipulating scientific data and knowledge structures. Here we consider the processes and activities in the same order as we discussed the structures that they generate in the previous subsection.

The process of observation involves inspecting the environmental setting by focusing an instrument, sometimes simply the agent's senses, on that setting. The result is a concrete description of the setting, expressed in terms from the agent's taxonomy and guided by the model of the setting. Since one can observe many things in any given situation, the observer must select some aspects to record and some to ignore.

As we have noted, scientific discovery is concerned with generating scientific knowledge in the form of taxonomies, laws and theories. These can be generated directly from observations (and possibly background knowledge), but, quite often, scientists modify an existing taxonomy, law, or theory to take into account anomalous observations that it cannot handle.

- *Taxonomy formation (and revision)* involves the organization of observations into classes and subclasses, along with the definition of those classes. This process may operate on, or take into account, an existing taxonomy or background knowledge. For instance, early chemists organized certain chemicals into the classes of acids, alkalis, and salts to summarize regularities in their taste and behavior. As time went on, they refined this taxonomy and modified the definitions of each class.
- *Inductive law formation (and revision)* involves the generation of empirical laws that cover observed data. The laws are stated using terms from the agent's taxonomy, and they are constrained by a model of the setting and possibly by the scientist's background knowledge. In some cases, the scientist may generate an entirely new law; in others, he may modify or extend an existing law.

– *Theory formation (and revision)* stands in the same relation to empirical laws as does law formation to data. Given one or more laws, this activity generates a theory from which one can derive the laws for a given model by explanation. Thus, a theory interconnects a set of laws into a unified account. Theory revision responds to anomalous phenomena or laws that cannot be explained by an existing theory, producing a revised theory that explains the anomaly while maintaining the ability to cover existing laws.

While some scientific activities revolve around inductive reasoning, others instead rely on deduction. Scientists typically derive predictions from laws or models, and sometimes they even deduce laws from theoretical principles.

– In contrast to inductive law discovery from observations, *deductive law formation* starts with a theory and uses an explanatory framework to deduce both a law and an explanation of how that law follows from the theory.
– The *prediction* process takes a law, along with a particular setting, and produces a prediction about what one will observe in the setting. Typically, a scientist derives a model from the law, taking into account the setting's particularities, and derives a prediction from the model. The analogous process of *postdiction* takes place in cases where the scientist must account for existing observations. Prediction and postdiction stand in the same relation to each other as deductive law formation and explanation.
– The process of *explanation* connects a theory to a law (or a law to a prediction) by specifying the deductive reasoning that derives the law from the theory. In the context of evaluation, a successful explanation lends support to the theory or law. If explanation fails, then an anomaly results that may trigger a revision of the theory or law. Explanation and deductive law formation are closely related, although explanation aims to account for a law that is already known. Also, in some fields explanation relies on abductive reasoning that leads the scientist to posit unobserved structures or processes, rather than deduction from given premises.

To assess the validity of theories or laws, scientists compare their predictions or postdictions with observations. This produces either consistent results or anomalies, which may serve to stimulate further theory or law formation or revision. This process is called *evaluation* and generally follows experimentation and observation.

Experimentation involves experimental design and manipulation. *Experimental design* specifies settings in which the scientist will collect measurements. Typically, he varies selected aspects of the environment (the independent variables) to determine their effect on other aspects (the dependent variables). He then constructs a physical setting (this is called *manipulation*) that corresponds to the desired environmental conditions and carries out the experiment.

Observation will typically follow or will be interleaved with systematic experimentation, in which case we call it active observation. However, there are fields and phenomena where experimental control is difficult, and sometimes

impossible. In such cases the scientist can still collect data to test his hypotheses through passive observation.

## 4 History of Research on Computational Discovery of Scientific Knowledge

### 4.1 A Brief Historical Account of Computational Scientific Discovery

Now that we have considered the goals of research on computational discovery and the elements it involves, we can provide some historical context for the work reported in this volume. The idea that one might automate the discovery of scientific knowledge has a long history, going back at least to the writings of Francis Bacon (1620) and John Stuart Mill (1900). However, the modern treatment of this task came from Herbert Simon, who proposed viewing scientific discovery as an instance of heuristic problem solving. In this paradigm, one uses mental operators to transform one knowledge state into another, invoking rules of thumb to select from applicable operators, choose among candidate states, and decide when one has found an acceptable solution. Newell et al. (1958) proposed this framework as both a theory of human problem solving and an approach to building computer programs with similar abilities.

Simon (1966) suggested that, despite the mystery normally attached to scientific discovery, one might explain it in similar terms. He noted that scientific theories can be viewed as knowledge states, and that mental operations can transform them in response to observations. He even outlined an approach to explaining creative phenomena such as scientific insight using these and other established psychological mechanisms. Simon's early papers on this topic only outlined an approach to modeling discovery as problem-space search, but they set a clear research agenda that is still being explored today.

The late 1970s saw two research efforts that transformed Simon's early proposals into running computer programs. The AM system (Lenat, 1978) rediscovered a variety of concepts and conjectures in number theory, starting from basic concepts and heuristics for combining them. The Bacon system (Langley, 1979; Langley et al., 1983) rediscovered a number of numeric laws from the history of physics and chemistry, starting from experimental data and heuristics for detecting regularities in them. Despite many differences, both systems utilized data-driven induction of descriptive laws and were demonstrated on historical examples. Together, they provided the first compelling evidence that computational scientific discovery was actually possible. There is no question that these early systems had many limitations, but they took the crucial first steps toward understanding the discovery process.

The following decade saw a number of research teams build on and extend the ideas developed in AM and Bacon. A volume edited by Shrager and Langley (1990) includes representative work from this period that had previously been scattered throughout the literature in different fields. This collection reported

work on discovery of descriptive laws, but it also included chapters on new topics, including the formation of explanatory models, hypothesis-driven experimental design, and model revision. On reading this book, one gets the general impression of an active research community exploring a variety of ideas that address different facets of the complex endeavor we know as science.

The early work on computational discovery focused on reconstructions from the history of science that were consistent with widely accepted theories of human cognition. This was an appropriate strategy, in that these examples let researchers test their methods on relatively simple problems for which answers were known, yet that were relevant because they had once been challenging to human scientists. Such evaluations were legitimate because it was quite possible to develop methods that failed on historical examples, and many approaches were ruled out in this manner. However, critics often argued that the evidence for computational discovery methods would be more compelling when they had uncovered new scientific knowledge rather than rediscovered existing results.

The period from 1990 to 2000 produced a number of novel results along these lines, a number of which have been reviewed by Valdez-Perez (1996) and Langley (2000). These successes have involved a variety of scientific disciplines, including astronomy, biology, chemistry, metallurgy, ecology, linguistics, and even mathematics, and they run the gamut of discovery tasks, including the formation of taxonomies, qualitative laws, numeric equations, structural models, and process explanations. What they hold in common is that each led to the discovery of new knowledge that was deemed significant enough to appear in the literature of the relevant field, which is the usual measure of scientific success. The same techniques have also proved successful in engineering disciplines, in which analogous modeling tasks also arise. These results provide clear evidence that our computational methods are capable of making new discoveries, and thus respond directly to early criticisms.

Another development during this period was the emergence of the data mining movement, which held its first major conference in 1995. This paradigm has emphasized the efficient induction of accurate predictive models from very large data sets. Typical applications involved records of commercial transactions, but some data-mining work has instead dealt with scientific domains. Although research in this area is sometimes referred to as "knowledge discovery" (Fayyad et al., 1996), the resulting models are generally encoded as decision trees, logical rules, Bayesian networks, or other formalisms invented by computer scientists. Thus, it contrasts with the smaller but older movement of computational scientific discovery, which focuses on knowledge cast in formalisms used by practicing scientists and which is less concerned with large data sets than with making the best use of available observations.

### 4.2 Lessons Learned for the Computational Discovery of Scientific Knowledge

Developments in both data mining and computational scientific discovery make it clear that technologies for knowledge discovery are mature enough for

application, but this does not mean there remains no need for additional research. In another paper, Langley (2002) recounts some lessons that have emerged from work in scientific domains, which we review here.

1. The output of a discovery system should be communicated easily to domain scientists. This issue deserves mention because traditional notations developed by machine learning researchers, such as decision trees or Bayesian networks, differ substantially from formalisms typical to the natural sciences, such as numeric equations and reaction pathways. Most work on computational scientific discovery attempts to generate knowledge in an established notation, but communicability is a significant enough issue that it merits special attention.
2. Discovery systems should take advantage of background knowledge to constrain their search. Most research in computational scientific discovery and data mining emphasizes the construction of knowledge from scratch, whereas human scientists often utilize their prior knowledge to make tasks tractable. For instance, science is an incremental process that involves the gradual improvement and extension of previous knowledge, which suggests the need for more work on methods for revising scientific laws, models, and theories. In addition, scientists often use theoretical constraints to guide their construction of models, so more work on this topic is needed as well.
3. Computational methods for scientific discovery should be able to infer knowledge from small data sets. Despite the rhetoric common in papers on data mining, scientific data are often rare and difficult to obtain. This suggests an increased focus on ways to reduce the variance of discovered models and mitigate the tendency to overfit the data, as opposed to developing methods for processing large data sets efficiently.
4. Discovery systems should produce models that move beyond description to provide explanations of data. Early work focused on discovery of descriptive regularities that summarized data, and most work on data mining retains this focus. However, mature sciences are generally concerned with explanatory accounts that incorporate theoretical variables, entities, or processes, and we increased work on methods that support such deeper scientific reasoning.
5. Computational discovery systems should support interaction with domain scientists. Most discovery research has focused on automated systems, yet few scientists want computers to replace them. Rather, they want computational tools that can assist them in constructing and revising their models. To this end, we need more work on interactive systems that let users play at least an equal role in the discovery process.

The chapters in this book respond directly to the first four of these issues, which suggests that they are now receiving the attention they deserve from researchers in the area. However, the fifth topic is not represented, and we hope it will become a more active topic in the future.

## 5 Overview of the Book

The chapters of the book present state-of-the-art approaches to computational scientific discovery, representing recent progress in the area. These approaches correspond to various scientific activities and deal with different scientific knowledge structures. Note, however, that the main focus of this edited volume is on inductive model formation from observed data. This is in contrast with a previous related book (Shrager & Langley, 1990) where most of the research presented concerned the formation and revision of scientific theories and laws.

In the first part of the book, titled "Equation Discovery & Dynamic Systems Identification", the focus is on establishing models of dynamic systems, i.e., systems that change their state over time. The models are mostly based on equations, in particular ordinary differential equations that represent a standard formalism for modeling dynamic systems in many engineering and scientific areas. This is in contrast to the bulk of previous research on equation discovery, which focuses on algebraic equations. The first two chapters by Stole, Easley, and Bradley present the PRET reasoning tool for nonlinear system identification, i.e., for solving the task of establishing equation-based models of dynamic systems. PRET integrates qualitative reasoning, numerical simulation, geometric reasoning, constraint reasoning, backward chaining, reasoning with abstraction levels, declarative meta-control, and truth maintenance to identify a proper model structure and its parameters for the modeling task at hand. Background knowledge for building models guides the reasoning engine. While the first chapter focuses mainly on general modeling knowledge that is valid in different scientific and engineering domains, the focus of the second chapter is on representing and use of knowledge specific to the domain of interest. The second chapter also presents PRET's heuristics for performing active observation of the modeled dynamic system.

The following chapter by Todorovski and Džeroski provides an overview of equation discovery approaches to inducing models of dynamic systems. Equation discovery deals with the task of automated discovery of quantitative laws, expressed in the form of equations, in collections of measured data. It has advanced greatly from the early stage, when the focus was on reconstructing well-known laws from scientific textbooks, and state-of-the-art approaches deal with establishing new laws and models from observed data. Among the most important recent research directions in this area has been the use of domain knowledge in addition to measured data in the equation discovery process. The chapter shows how modeling knowledge specific to the domain at hand can be integrated in the process of equation discovery for establishing and revising comprehensible models of real-world dynamic systems.

The chapter by Washio and Motoda also presents an approach to formulating equation-based models and laws from observed data. They use results from measurement theory (in particular the Buckingam theorem) about how to properly combine variables measured using different measurement units and scales. These rules are used to constrain the space of candidate models and laws for the observed phenomena. The second part of the chapter discusses the conditions that equations have to satisfy in order to be considered communicable knowledge.

The next two chapters deal with establishing models from Earth science data. The first chapter by Saito and Langley presents an approach to revising existing scientific models cast as sets of equations. The revision is guided by the goal of reducing the model error on newly acquired data and allows for revising parameter values, intrinsic properties, and functional forms used in the model equations. The second chapter by Schwabacher et al. shows how standard machine learning methods can be used to induce models that are represented in formalisms specific to the scientific fields of artificial intelligence and machine learning and yet understandable and communicable to Earth scientists.

In the next chapter, Colton reviews research on computational discovery in pure mathematics, where the focus is on theory and law formation. The author puts special emphasis on his own work in the area of taxonomy formation in mathematics, especially with respect to identifying important classes of numbers.

The last chapter in the first part of the book by Zhao et al. presents a spatial aggregation method for identifying spatio-temporal objects in observations. The method recursively aggregates data into objects and artifacts at higher levels of abstraction. Although the presented method does not correspond directly to any of the scientific activities presented in this introduction, it can be a very useful tool for aiding the processes of taxonomy, law, and model formation.

While the first part of the book focuses on a class of methods and covers a variety of scientific fields and areas, the focus of the second part is on computational scientific discovery in biomedicine and bioinformatics. The first three chapters are in line with the first part of the book and continue with the theme of model formation. However, the model representation formalisms change from equations to formalisms specific to biomedicine, such as chemical reaction networks and genetic pathways.

The chapter by Koza et al. deals with the problem of inducing chemical reaction networks from observations of compounds concentration through time. The authors show that chemical reaction networks can be transformed to (systems of) ordinary differential equations. They present and evaluate a genetic programming approach to inducing a restricted class of equations that correspond to chemical reaction networks.

The chapter by Zupan et al. presents a reasoning system for inferring genetic networks, i.e., networks of gene influences on one another and on biological outcomes of interest. The system uses abduction and qualitative simulation to transform observations into constraints that have to satisfied by a network that would describe observed experimental data best. The following chapter by Garrett et al. also represents genetic networks as qualitative models and uses qualitative simulation to match them against observed data. The authors present and evaluate a method for inducing qualitative models from observational data that is based on inductive logic programming.

The chapter by King et al. deals with the application of inductive logic programming methods to the task of analyzing a complex bioinformatic database in the domain of functional genomics. The authors discuss the importance of integrating background knowledge in the process of scientific data analysis and show

that inductive logic programming tools provide an appropriate environment for the integration of knowledge and data in the process of scientific discovery. The work presented in the chapter is the initial step that later lead to the development of a robot scientist, capable of automatically performing a variety of scientific activities. The robot scientist project is one of the most exciting recent developments in the field of computational scientific discovery (King et al., 2004).

Finally, the last two chapters present approaches to forming hypotheses by connecting disconnected scientific literatures on the same topic. Weber presents a general model that, based on connections already published in the scientific literature between a symptom and a disease on one hand and connections between an active substance (chemical compound) and a symptom on the other hand, establishes a hypothesis that the chemical compound can be used for treatment of the disease. The hypothesis is of interest, if the relation between the disease and the compound has not been established before while evidences for the other two relations are well presented in scientific literature. In the final chapter, Hristovski et al. present an interactive system for literature discovery and apply it to the task of identifying gene markers for a particular disease. The system uses association rule mining to find relations between medical concepts from a bibliographic database and uses them to discover new relations that have not been reported in the medical literature yet.

## References

Achinstein, P. (ed.): Science rules: A historical introduction to scientific methods. The Johns Hopkins University Press, Baltimore (2004)

Bacon, F.: The new organon and related writings. Liberal Arts Press, New York (1620/1960)

Darden, L.: Reasoning in biological discoveries. Cambridge University Press, Cambridge (2006)

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery in databases. AI Magazine 17, 37–54 (1996)

Gower, B.: Scientific method. Routledge, Florence, KY (1996)

Hanson, N.R.: Patterns of discovery. Cambridge University Press, Cambridge (1958)

King, R.D., Whelan, K.E., Jones, F.M., Reiser, P.G.K., Bryant, C.H., Muggleton, S., Kell, D.B., Oliver, S.G.: Functional genomic hypothesis generation and experimentation by a robot scientist. Nature 427, 247–252 (2004)

Klemke, E.D., Hollinger, R., Rudge, D.W., Kline, A.D. (eds.): Introductory readings in the philosophy of science. 3rd edn. Prometheus Books, Amherst, NY (1998)

Lakatos, I.: Proofs and refutations: The logic of mathematical discovery. Cambridge University Press, Cambridge (1976)

Langley, P.: Rediscovering physics with Bacon.3. In: Proceedings of the Sixth International Joint Conference on Artificial Intelligence, pp. 505–507. Morgan Kaufmann, Tokyo (1979)

Langley, P.: The computational support of scientific discovery. International Journal of Human-Computer Studies 53, 393–410 (2000)

Langley, P.: Lessons for the computational discovery of scientific knowledge. In: Proceedings of First International Workshop on Data Mining Lessons Learned, pp. 9–12. University of New South Wales, Sydney (2002)

Langley, P., Bradshaw, G.L., Simon, H.A.: Rediscovering chemistry with the bacon system. In: Michalski, R.S., Carbonell, J.G., Mitchell, T.M. (eds.) Machine learning: An artificial intelligence approach, Morgan Kaufmann, San Mateo (1983)

Lenat, D.B.: The ubiquity of discovery. Artificial Intelligence 9, 257–285 (1978)

Merriam-Webster.: Merriam-webster's collegiate dictionary. 11th edn. Merriam-Webster, Springfield, MA (2003)

Mill, J.S.: A system of logic ratiocinative and inductive being a connected view of the principles of evidence and the methods of scientific investigation. 8th edn. Longmans, Green, & Co., London (1900)

Newell, A., Shaw, J.C., Simon, H.A.: Chess-playing programs and the problem of complexity. IBM Journal of Research and Development 2, 320–325 (1958)

Popper, K.R.: The logic of scientific discovery. Hutchinson, London (1959)

Shrager, J., Langley, P. (eds.): Computational models of scientific discovery and theory formation. Morgan Kaufmann, San Mateo (1990)

Simon, H.A.: Scientific discovery and the psychology of human problem solving. In: Colodny, R.G. (ed.) Mind and cosmos: Essays in contemporary science and philosophy, University of Pittsburgh Press, Pittsburgh (1966)

Valdez-Perez, R.E.: Computer science research on scientific discovery. Knowledge Engineering Review 11, 51–66 (1996)