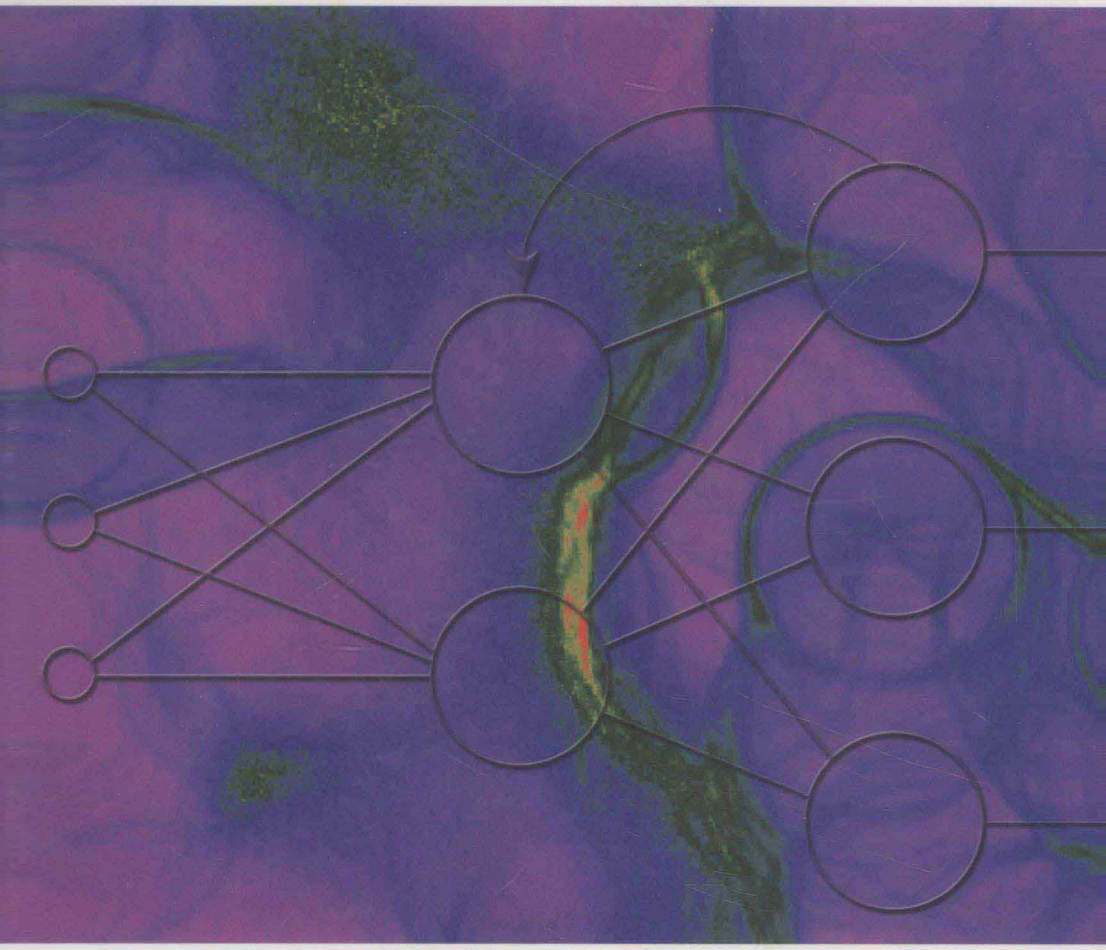
USING ARTIFICIAL INTELLIGENCE IN CHEMISTRY AND BIOLOGY

A Practical Guide



Hugh Cartwright

CRC Press
Taylor & Francis Group

USING ARTIFICIAL INTELLIGENCE IN CHEMISTRY AND BIOLOGY

A Practical Guide

Hugh Cartwright

Chapter 10, *Evolvable Developmental Systems*, contributed by

Nawwaf Kharma
Associate Professor, Department of Electrical and Computer Engineering,
Concordia University, Montreal, Canada

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-13: 978-0-8493-8412-7 (Hardcover)

Library of Congress Cataloging-in-Publication Data

Cartwright, Hugh M., 1948-
Using artificial intelligence in chemistry and biology : a practical guide / Hugh Cartwright ; chapter 10, Evolvable developmental systems, contributed by Nawwaf Kharma.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-8493-8412-7 (acid-free paper) 1. Chemistry--Data processing. 2. Biology--Data processing. 3. Artificial intelligence. I. Kharma, Nawwaf. II. Title.

QD39.3.E46C375 2008
540.285'63--dc22 2007047462

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Preface

The use of Artificial Intelligence within science is growing at a remarkable rate. In the early 1990s, little was known about how Artificial Intelligence could be applied in a practical way to the physical and life sciences. At that stage, few experimental scientists showed any interest in the area. Now, hundreds of research papers are published each year and the numbers are rising rapidly. The change has been dramatic and yet, despite the growth, the field is still very young.

The upsurge of interest owes much to an increasing understanding of the scientific potential of Artificial Intelligence. This book explains in a lucid and straightforward way how these methods are used by scientists and what we can accomplish with them. Recognizing that not all experimental scientists are computer experts, the approach adopted here assumes no prior knowledge of Artificial Intelligence and no unusual skills in computer science or programming. It does, however, presume some scientific background. Each chapter is designed to take the reader quickly to the point at which meaningful scientific applications can be investigated.

Computer scientists may use this book to gain a clearer picture of how experimental scientists use Artificial Intelligence tools. Chemists, biochemists, physicists, and others in the experimental sciences who have data to analyze or simulations to run will find tools within these pages that may speed up their work or make it more effective. For both groups, the aim of this book is to encourage a broader application of these methods.

Many people have contributed to the production of this book. The final chapter was written by Nawwaf Kharma, following several months on sabbatical at Oxford University. His perceptive and challenging chapter gives a glimpse of the future of Artificial Intelligence in science. EJS, a software tool for the construction of simulations in Java, has been used to perform a number of calculations for this book. A complete version of EJS is included on the accompanying CD with the permission of Francisco Esquembre, to whom I am most grateful. Most of the figures in this text have been prepared by John Freeman, adding both clarity and humor to it. The editorial staff at Taylor & Francis Group, particularly Lance Wobus, Russ Heaps, and Pat Roberson, have been a regular source of encouragement and expert advice. Numerous members of my research group, past and present, have contributed to the ideas that are crystallized here, notably "the two Alexes,"whose comments about the interface between Artificial Intelligence and science are often thought-provoking. Finally, my wife, Susie, tolerating my long and unsociable working hours, and my daughter, Jenny, a distant but engaging e-mail voice, have provided less tangible but invaluable contributions. I am grateful to them all.

Errors are almost unavoidable in a text of this sort and I shall be grateful to be notified of any that may be spotted by readers.

Hugh Cartwright

Chemistry Department, Oxford University
St Anne's College and Oriel College, Oxford, United Kingdom

The Author

Hugh M. Cartwright is a member of the Chemistry Department at Oxford University in the United Kingdom, where he is Laboratory Officer. He is also a Supernumerary Fellow at St. Anne's College, Oxford, and a Lecturer in Physical Chemistry at Oriel College, Oxford. His research interests lie in the broad field of the application of computational methods from Artificial Intelligence to problems within science. Past and current research, carried out by some seventy students, has covered a wide variety of areas. Typical of these are studies of biodegradation, the prediction of the properties of polymers, the analysis of oil spill data, automatic elucidation of scientific rules, the design of organic synthetic routes, the optimization of solid-state phosphor composition, quantitative structure-activity relationships (QSARs), unraveling mass spectrometric data from crude oils, and the analysis of large biomedical datasets. The work of his group focuses primarily on the use of genetic algorithms, neural networks, self-organizing maps, and growing cell structures.

He has written, edited, and provided invited chapters in a number of textbooks on the use of Artificial Intelligence in science and continues to run an active research group in the area. He has taught at universities in Canada, the United States, Ireland, and the United Kingdom, and is a board member of the CoLoS (Conceptual Learning of Science) International Consortium.

Contents

1

Artificial Intelligence

Computers don't understand. At least, they don't understand in the way that we do. Of course, there are some things that are almost beyond comprehension: Why do women collect shoes? What are the rules of cricket? Why do Americans enjoy baseball? And why are lawyers paid so much?

These are tough questions. As humans, we can attempt to answer them, but (at least for the next few years) computers cannot help us out. Humans possess far more "intelligence" than computers, so it is to be expected that if a human struggles to understand shoe collecting, a computer will be totally baffled by it. This makes it all the more surprising that computers that use "artificial intelligence" can solve a variety of scientific tasks much more effectively than a human.

Artificial Intelligence (AI) tools are problem-solving algorithms. This book provides an introduction to the wide range of methods within this area that are being developed as scientific tools. The application of AI methods in science is a young field, hardly out of diapers in fact. However, its potential is huge. As a result, the popularity of these methods among physical and life scientists is increasing rapidly.

In the conventional image of AI, we might expect to see scientists building laboratory robots or developing programs to translate scientific papers from Russian to English or from English to Russian, but this does not accurately reflect the current use of AI in science. The creation of autonomous robots and the perfection of automatic translators are among the more important areas of research among computer scientists, but the areas in which most experimental scientists are involved are very different.

Science is dominated by problems; indeed, without them there would be hardly any science. When tackling a new problem, scientists are generally less concerned about *how* they solve it, provided that they get a solution, than they are about the quality of the solution that is obtained. Thus, the pool of mathematical and logical algorithms that scientists use for problem solving continues to grow. That pool is now being augmented by AI methods.

Until the second half of the 1990s, experimental scientists, by and large, knew little about AI, suspecting (without much evidence) that it might be largely irrelevant in practical science. This turned out to be an unduly pessimistic view. AI algorithms are in reality widely applicable and the power of these methods, allied to their simplicity, makes them an attractive proposition in the analysis of scientific data and for scientific simulation. The purpose of this book is to introduce these methods to those who need to solve scientific problems, but do not (yet) know enough about AI to take advantage

of it. The text includes sufficient detail that those who have some modest programming skills, but no prior contact with AI, can rapidly set about using the techniques.

1.1 What Is Artificial Intelligence?

Computer scientists usually define AI in terms of what it can accomplish. They are especially interested in the creation of intelligent software that can reproduce the kind of behavior that humans would recognize as intelligent, such as understanding language or conducting a conversation.

Some of this work on human-like behavior spills over into the scientific laboratory where work progresses on producing intelligent robotic systems for chemical or biological laboratories. Robotic sample handlers are commonplace in science, and already incorporate enough rudimentary intelligence to be able to cope with the demands of the job. However, a robotic sampler is really pretty dumb and, in truth, that is how it should be. A sampler whose role is to feed samples into a mass spectrometer should do no more than unquestioningly follow predefined rules of priority. If it had ideas above its station and started making its own decisions about whose sample would be processed next, based on the color of the users' hair or the number of syllables in their name perhaps, it would soon be headed for the great sampler scrap heap in the sky.

Experimental scientists tend not to be interested in robotic samplers, even those with attitude; instead they want software that will solve problems. For them, AI tools may be of considerable value as problem solvers. In this book, we explore such uses.

For most scientific purposes, a reasonable working definition of AI is: "A computer program that can learn."

Although computer scientists might quibble at this definition, it describes most AI methods rather well. Methods that depend on learning, such as artificial neural networks and genetic algorithms, are already proving their worth in science, and there are others, such as growing cell structure networks, whose potential is just starting to be recognized.

The learning that AI algorithms depend on can be achieved in several different ways. An expert system (Chapter 7) is tutored by a human expert and starts off resembling an empty box awaiting information. This requires that a scientist spend long periods feeding relevant information into the system. By contrast, artificial neural networks (Chapter 2) inspect examples of what they are required to understand and learn by extracting examples from a database or by interacting with a simulation. Classifier Systems (Chapter 9) talk directly with their environment, attempting to control it, receiving feedback from it, and learning from any mistakes. The self-organizing map (Chapter 3) makes its deductions by just looking at data without having to be told what it is expected to learn.

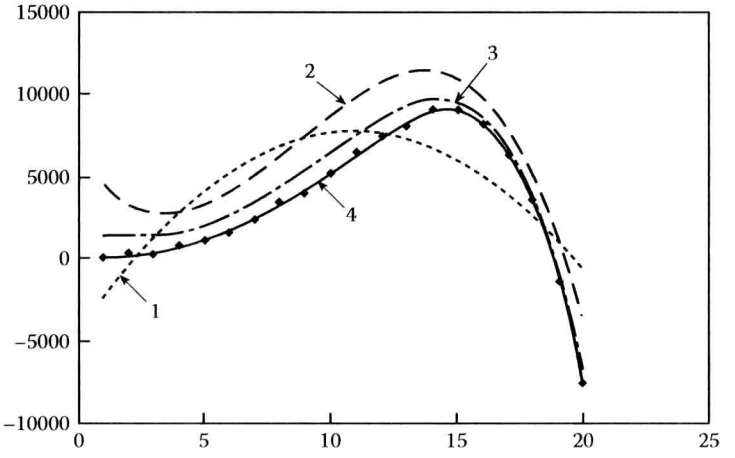


FIGURE 1.1
Progressive improvements in the attempts by a genetic algorithm to find a curve that fits a set of data points well (lines 1 to 3). The solid line is fitted using standard least squares.

But, you may have spotted a problem here. If an AI program has to learn, does that not suggest that its performance might be fairly hopeless when we first ask it to solve a problem? Indeed it does, as Figure 1.1 indicates.

The figure shows a polynomial fit to a dataset calculated according to a standard least squares algorithm (solid line); this is compared with a series of attempts to find a fit to the same data using a genetic algorithm.

The genetic algorithm is unimpressive, at least to begin with (line 1). The quality of the fit that it finds does improve (lines 2 and 3) and eventually it will reach a solution that matches the least squares fit, but the algorithm takes far longer to find the solution than standard methods do, and the fit is no better.

1.2 Why Do We Need Artificial Intelligence and What Can We Do with It?

So why bother? If established and reliable methods exist to solve scientific problems, surely there is no point in using methods from AI that may be slower or less effective? The key here is the phrase: "if established and reliable methods exist." In many situations, there is no need to turn to AI methods, but not every scientific problem is as straightforward as finding the line of best fit. When more conventional methods are unavailable, or are of insufficient power, an AI method may be just what is needed.

There are several reasons why it may be difficult to find an adequate solution to a scientific problem; the two most common are scale and complexity.

In many types of problems, the number of possible solutions can be huge. The folding of a protein is a typical example. Proteins fold to take up a conformation of minimum energy, although the conformation that the protein adopts in a living organism may not be that of the global minimum energy. The number of ways in which a large molecule, such as a protein, can be folded so that its energy is at least locally minimized and all bond lengths and angles are reasonable is so large that it is impossible to investigate all of them computationally. As it is not feasible to check every conformation, some search method is needed that makes intelligent decisions about which conformations to assess and which to ignore.

A problem of similar scale is the identification of useful drugs from among the huge number that could, in principle, be prepared. The number of molecules that might potentially act as drugs can only be estimated, but it is probably of the order of 10^{40}. Of this huge number of possibilities only a tiny fraction will ever be synthesized, and of those only a handful will make it through the process of high throughput screening to human trials. Any computational procedure that is devised to screen molecules for possible value as drugs must be able, somehow, to search through this huge number of candidates to identify promising structures without actually inspecting every molecule (Figure 1.2).

The chemical flowshop, which we shall meet in Chapter 5, is similarly challenging to solve. Many different chemicals are synthesized within a single line of processing units in a defined sequence. The efficiency of the flowshop, which is determined by the length of time required to synthesize the complete group of chemicals, is strongly correlated with the order in which chemicals are made. Therefore, determining a near-optimum order is important in the operation of flowshops if the process is to be financially viable. For n chemicals, the number of possible sequences is $n!$, a number

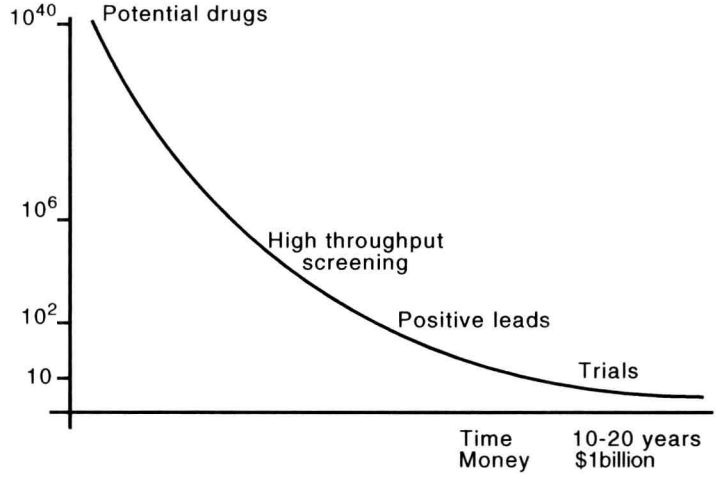


FIGURE 1.2
The drug development process.