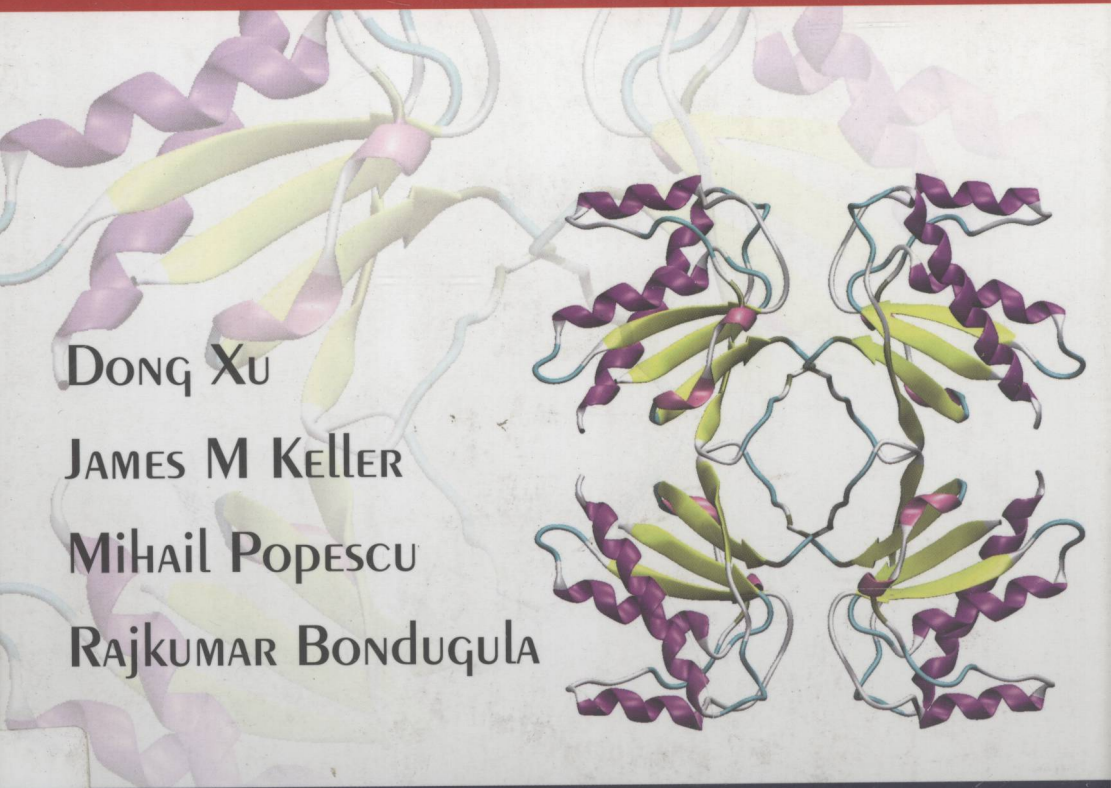
Series on Advances in Bioinformatics and Computational Biology – Volume 9

# APPLICATIONS OF FUZZY LOGIC IN BIOINFORMATICS



Dong Xu
James M Keller
Mihail Popescu
Rajkumar Bondugula

Imperial College Press

Series on Advances in Bioinformatics and Computational Biology – Volume 9



# APPLICATIONS OF FUZZY LOGIC IN BIOINFORMATICS

Dong Xu
James M Keller
Mihail Popescu
Rajkumar Bondugula

University of Missouri-Columbia, USA



Imperial College Press

ICP

*Published by*

Imperial College Press
57 Shelton Street
Covent Garden
London WC2H 9HE

*Distributed by*

World Scientific Publishing Co. Pte. Ltd.
5 Toh Tuck Link, Singapore 596224
*USA office:* 27 Warren Street, Suite 401-402, Hackensack, NJ 07601
*UK office:* 57 Shelton Street, Covent Garden, London WC2H 9HE

**British Library Cataloguing-in-Publication Data**
A catalogue record for this book is available from the British Library.

**APPLICATIONS OF FUZZY LOGIC IN BIOINFORMATICS**
**Series on Advances in Bioinformatics and Computational Biology — Vol. 9**



ISBN-13 978-1-84816-258-7
ISBN-10 1-84816-258-8

Printed in Singapore by World Scientific Printers

# APPLICATIONS OF FUZZY LOGIC IN BIOINFORMATICS

## SERIES ON ADVANCES IN BIOINFORMATICS AND COMPUTATIONAL BIOLOGY



**ISSN: 1751-6404**

---

Vol. 1: Proceedings of the 3rd Asia-Pacific Bioinformatics Conference
*Eds: Yi-Ping Phoebe Chen and Limsoon Wong*

Vol. 2: Information Processing and Living Systems
*Eds: Vladimir B. Bajic and Tan Tin Wee*

Vol. 3: Proceedings of the 4th Asia-Pacific Bioinformatics Conference
*Eds: Tao Jiang, Ueng-Cheng Yang, Yi-Ping Phoebe Chen and Limsoon Wong*

Vol. 4: Computational Systems Bioinformatics 2006
*Eds: Peter Markstein and Ying Xu*
ISSN: 1762-7791

Vol. 5: Proceedings of the 5th Asia-Pacific Bioinformatics Conference
*Eds: David Sankoff, Lusheng Wang and Francis Chin*

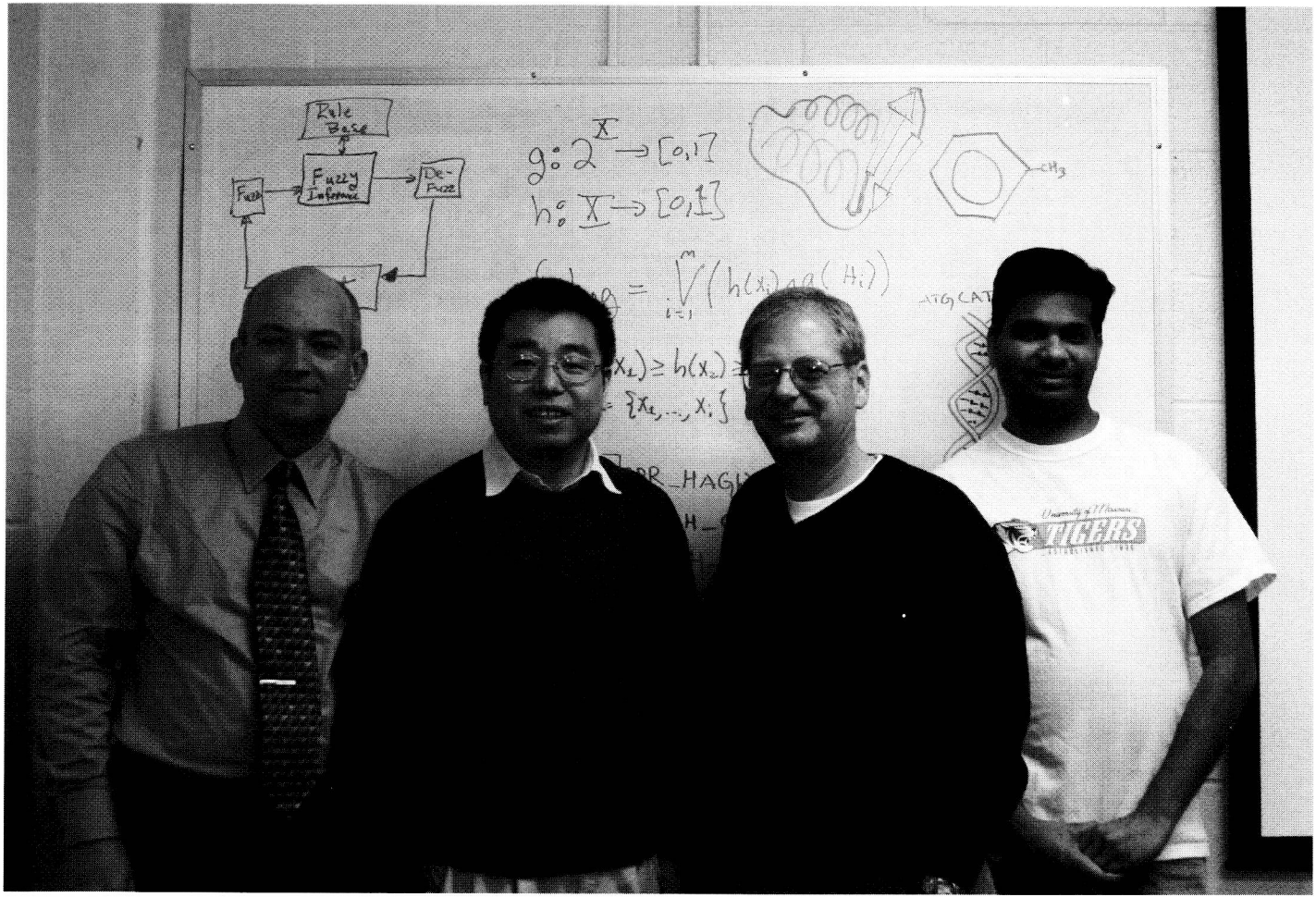
Vol. 6: Proceedings of the 6th Asia-Pacific Bioinformatics Conference
*Eds: Alvis Brazma, Satoru Miyano and Tatsuya Akutsu*

Vol. 7: Computational Methods for Understanding Bacterial and Archaeal Genomes
*Eds: Ying Xu and J. Peter Gogarten*

Vol. 8 Regulatory Genomics
*Eds: Leong Hon Wai, Sung Wing-Kin and Eleazar Eskin*

Vol. 9 Applications of Fuzzy Logic in Bioinformatics
*Eds: Dong Xu, James M. Keller, Mihail Popescu and Rajkumar Bondugula*

To our families




(*left to right*) Mihai, Dong, Jim, and Raj

# Foreword

Bioinformatics is one of the youngest and most exciting fields in modern science. During the past decade, bioinformatics has become a challenging arena of applications of a wide variety of concepts and sophisticated techniques drawn from mathematics, computer science and probability theory. Within the fuzzy logic community, the meteoric ascent of bioinformatics has led to a contentious question: Can fuzzy logic make a substantive contribution to advancement of bioinformatics? The pioneering work "Applications of Fuzzy Logic to Bioinformatics," co-authored by Professors Dong Xu, James Keller, Mihail Popescu and Dr. Rajkumar Bondugula, may be viewed as a persuasive argument in support of an affirmative answer to the question. The core argument is that fuzzy logic is needed to solve problems in bioinformatics which are beyond the reach of existing techniques. It should be noted that "Applications of Fuzzy Logic to Bioinformatics," is the first book on this subject.

Today, fuzzy logic plays a relatively minor role in the armamentarium of bioinformatics. A metric is the number of publications with "fuzzy" in title or abstract—publications which are listed in the PubMed database. The current rate is 300-400 papers per year. Will the same be true in a few years from now? My belief is that in coming years there will be a rapid growth in the visibility and importance of fuzzy-logic-based techniques in the literature of bioinformatics and, more generally, in the literature of biological and medical sciences. However, my belief is based not on a detailed familiarity with bioinformatics—a familiarity which I do not have—but on my understanding of what fuzzy logic has to offer.

There are many misconceptions about fuzzy logic. Fuzzy logic is not fuzzy. Basically, fuzzy logic is a precise logic of imprecision. The principal objective of fuzzy logic is formalization/mechanization of imprecision, uncertainty, incompleteness of information and partiality of truth. Bioinformatics data fit this description, in addition to having a huge mass and high dimensionality.

Science deals not with reality but with models of reality. Concomitantly, scientific progress is driven by a quest for better models of reality. In bioinformatics, modeling is focused on genes, genomes and related biological entities. Brilliant successes have been achieved through the use of models based on bivalent logic and probability theory. However, there are many problems, such as those discussed in "Applications of Fuzzy Logic to Bioinformatics," in which better results can be achieved with better models based on the use of fuzzy logic. What is widely unrecognized is that modeling techniques based on bivalent logic and probability theory are intrinsically less powerful than techniques which are based on fuzzy logic and fuzzy-logic-based probability theory. An important contribution of "Applications of Fuzzy Logic to Bioinformatics" is making the bioinformatics community aware of the powerful modeling capability of fuzzy logic.

The superior capability of fuzzy logic as a modeling language is one of the principal rationales for its use in bioinformatics and, more generally, in scientific theories. An elaboration of this assertion is in order.

In a general setting, let M(S) be a model of S. There are two basic metrics which might be associated with M(S). First, the goodness of M(S) as a model of S, call it cointension; and second, the computational complexity of M(S). In general, cointension and computational complexity are covariant in the sense that an increase in cointension of M(S) results in an increase in the computational complexity of M(S). Bivalent logic and probability theory are, respectively, special cases of fuzzy logic and fuzzy-logic-based probability theory. What this implies is that, viewed as a modeling language, bivalent logic and probability theory have an intrinsically lower power of cointension than fuzzy logic and fuzzy-logic-based probability theory. However, the reverse is true so far as computational complexity is concerned. What gives fuzzy logic an

advantage is that an increase in the computational complexity is far less important than an increase in cointension. This, in principle, is one of the main rationales for the use of fuzzy logic in bioinformatics. It should be noted that the relation between bivalent logic and fuzzy logic is similar in spirit to the relation between linear system theory and nonlinear system theory.

"Applications of Fuzzy Logic to Bioinformatics," serves three major purposes. First, it introduces fuzzy logic to the bioinformatics community. Second, it introduces bioinformatics to the fuzzy logic community; and third, it demonstrates that fuzzy logic has much to contribute to the advancement of bioinformatics. Professors James Keller, Dong Xu, Mihail Popescu and Dr. Rajkumar Bondugula, and the Imperial College Press deserve our thanks and congratulations for producing a work whose importance is hard to exaggerate. They deserve a loud applause.

*Lotfi A. Zadeh*
*Berkeley, CA*
*September 24, 2007*

# Preface

Science is entering a new era thanks to the Human Genome Project, one of the largest programs in molecular biology. This project was devoted to the sequencing of human DNA fragments, i.e., to the determination of the order of nucleic acids therein. These sequences represent the blueprint of life. Since the 1980s, the advent of the Human Genome Project and other DNA sequencing projects has led to exponential growth in molecular data. Genomic sequencing has opened a new avenue to study biological systems on large scales, paving the way for investigating other high-throughput data. Today, due to the availability of high-throughput measurement technologies, it is possible to use a broad range of experimental data to expand the genome-scale studies from biological sequences and protein structures to higher-level functions and phenotypes. For example, microarray technology is a powerful tool to systematically measure gene expression across whole cells and tissues under varying experimental conditions or over a time course. As massive data are being generated, there is a strong demand for bioinformatics in data management, visualization, integration, analysis, modeling, and prediction. Bioinformatics has been developed extremely fast and has brought enormous impact to the research of biology and medicine in recent years. Thousands of bioinformatics databases and tools have been developed. More and more experimental biologists have realized the importance of bioinformatics, as the need for managing and analyzing the massive amount of data is evident. Many biologists now use bioinformatics tools themselves, especially through a Web interface.

As massive biological data have become a fundamentally important resource in biomedical sciences, researchers have developed various

bioinformatics algorithms and software tools to identify meaningful information (or statistically significant patterns) from data and correlate such information for discovery of new knowledge or prediction of biological properties. However, such tasks are often highly challenging. The information-rich data are heterogeneous and ambiguous in nature. They are often noisy and incomplete, as well as containing misleading outliers. Furthermore, biological systems, due to adaptability, evolution, redundancy, robustness, and emergence, are extremely complex. The challenge has drawn a wide range of studies from computer sciences, and various computer science technologies have been applied. The most notable applications include dynamic programming, neural networks, hidden Markov models, support vector machines, etc. Fuzzy set theory and fuzzy logic have also been used in bioinformatics, and we believe there is a much greater potential for their applications in bioinformatics in the future.

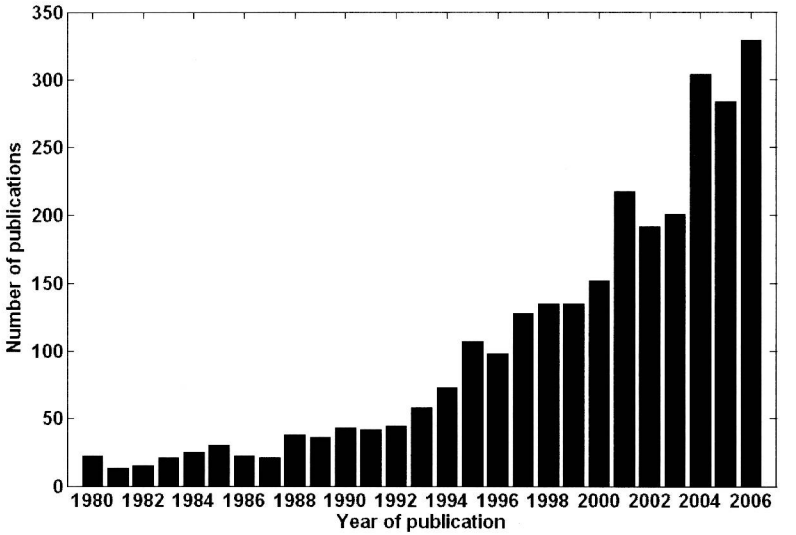



Figure 0.1. Number of publications containing the word "fuzzy" in their PubMed records since 1980.

Many biological systems and objects are intrinsically fuzzy as their properties and behaviors contain randomness or uncertainty. In addition,

it has been shown that exact or optimal methods have significant limitations in many bioinformatics problems. Fuzzy set theory and fuzzy logic are ideal to describe some biological systems/objects and provide good tools for many bioinformatics problems. The applications of fuzzy concepts and approaches have been growing at an exponential rate. Figure 0.1 illustrates the number of publications that contain the word "fuzzy" in their titles or abstracts in PubMed, a literature database mainly for biomedical research. Currently the number is increasing at a rate of about one publication per day. While a number of books have been published covering applications of other computational intelligence techniques in bioinformatics, no book addresses the applications of fuzzy set theory and fuzzy logic in bioinformatics. As researchers in this area and educators at a university, we feel that there is an urgent need for a comprehensive and systematic book covering this topic. Hence, this is our motivation in writing this book.

We developed the text in a way that is useful to a broad readership, including students, postdoctoral fellows, and senior investigators moving into the field, as well as professional practitioners/bioinformatics experts. We expect that the book can be used as a textbook for upper undergraduate-level or graduate-level bioinformatics courses. Bioinformatics applications using fuzzy set theory or fuzzy logic often require good understanding of the biological background and the computational algorithms. In our case, no prerequisite in biology is needed, and only college-level calculus is required, for reading this book. In other words, a dedicated reader with a college degree in computational, biological or physical science should be able to follow the book without much difficulty. To facilitate learning and to maximize the benefit of the book, we provide a comprehensive introduction in fuzzy set theory and an appendix in basic biological concepts. We also wish to promote more research in applying fuzzy approaches in bioinformatics through this book, especially to provide an informative source for beginners entering bioinformatics as young students or as experienced researchers coming from other disciplines.

In this book, we discuss why and how fuzzy concepts and methods can play an important role in studying biological problems. We have designed the chapters to comprehensively address several important

bioinformatics topics using fuzzy concepts and approaches. In addition, chapters have been connected seamlessly through a systematic design of the overall structure of the book. We start with an introduction to bioinformatics and then introduce fundamentals of fuzzy set theory and fuzzy logic. We focus on three examples (measurement of ontological similarity, protein structure prediction/analysis, and microarray data analysis). We also review other bioinformatics applications using fuzzy techniques. Finally we summarize and provide a future outlook. Furthermore we provide two appendices, one on fundamental biological concepts and one on online resources related to the book.

Chapter 1 (Introduction to Bioinformatics) discusses the scope of bioinformatics, including biological sequence analysis, protein structure analysis and prediction, gene expression data analysis, computational proteomics, gene ontology and biological pathway prediction. We will illustrate what the challenges in the fields are and why fuzzy logic can help.

Chapter 2 (Introduction to Fuzzy Set Theory and Fuzzy Logic) introduces fuzzy set theory and fuzzy logic. We will review the history of the field (together with types of successful applications). We will explain the key concepts and major methods, including fuzzy memberships, fuzzy clustering, fuzzy inference, etc. (tailored to potential bioinformatics applications).

Chapter 3 (Fuzzy Similarities in Ontologies) reviews some of the measures that can be used to compute the similarity between gene products annotated with terms from an ontology. We will introduce new fuzzy measures for computing ontological similarity between genes that avoid the problems of the traditional measures and, in addition, can account for information uncertainty. We will present several applications of the fuzzy similarity measures such as gene clustering and gene function summarization using the Gene Ontology terms. At the end of the chapter, we will present the application of the ontological similarity to computational intelligence algorithms such as fuzzy rule systems.

Chapter 4 (Fuzzy Logic in Structural Bioinformatics) introduces application of fuzzy logic in protein secondary structure prediction,

protein solvent accessibility prediction, and protein structure comparison/classification. We will show our computational results and describe related computational tools.

Chapter 5 (Application of Fuzzy Logic in Microarray Data Analyses) provides a review of several microarray processing algorithms for gene selection and patient classification. We will then describe several clustering algorithms such as fuzzy c-means, relational fuzzy c-means and fuzzy co-clustering, and their use for gene selection.

Chapter 6 (Other Applications) reviews other types of bioinformatics applications using fuzzy set theory and fuzzy logic in the literature, including biological sequence motif identification, protein sequence alignment, protein subcellular localization prediction, 3D protein structure comparison, and computational proteomics.

Chapter 7 (Summary and Outlook) summarizes the whole book. We will discuss the advantages and limitations of using fuzzy set theory and fuzzy logic in bioinformatics. We will also provide an outlook of future applications and directions in using the fuzzy concept in molecular biology. Further related readings will be suggested.

Appendix I (Fundamental Biological Concepts) introduces some fundamental biological concepts for readers without a biological background. We will cover major biological subjects discussed in the book.

Appendix II (Online Resources) describes some of the free online resources, including tools, databases, and tutorials related to molecular biology, bioinformatics, and fuzzy set theory.

During the writing of this book, we have received help and support from our friends, colleagues, and families, to whom we wish to take this opportunity to express our deep gratitude and appreciation. First we would like to thank Imperial College Press, who contacted us to start this book project. During the writing of this book, Ms. Lenore Betts and Ms. Katie Lydon, editors at Imperial College Press, answered many of our questions and we are grateful their help. We like to thank Gerald L. Arthur, Tim Havens, Tran Hong Nha Nguyen, Yangjiong Su, Anders Wallqvist, and Jingfen Zhang for critically reviewing the drafts of the book and providing many helpful suggestions. We also want to thank our

families for their constant support and encouragement over about a year of intensive writing.

*Dong Xu*
*James Keller*
*Mihail Popescu*
*Rajkumar Bondugula*

# Contents