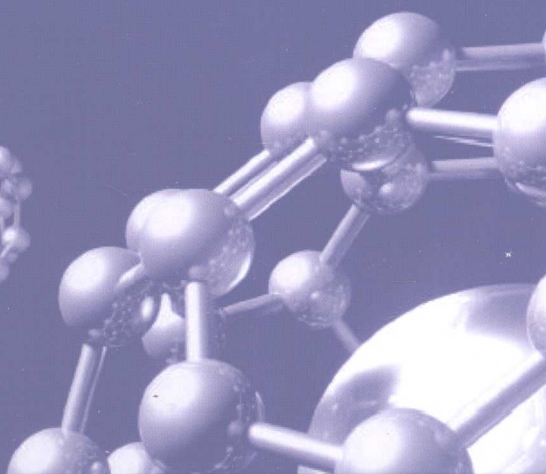
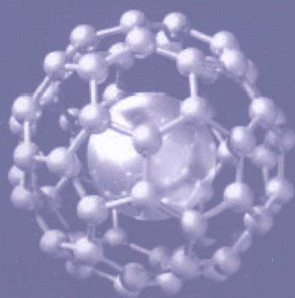
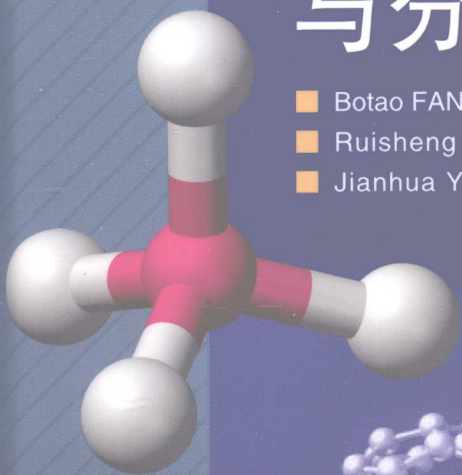


Computer Chemistry and Molecular Design

计算机化学 与分子设计

■ Botao FAN 范波涛
■ Ruisheng ZHANG 张瑞生
■ Jianhua YAO 姚建华



高等教育出版社
Higher Education Press

Computer Chemistry and Molecular Design

Botao FAN, Ruisheng ZHANG, Jianhua YAO

Higher Education Press

图书在版编目 (CIP) 数据

计算机化学与分子设计=Computer Chemistry and Molecular Design:
英文/范波涛,张瑞生,姚建华著. —北京:高等教育出版社,2009.3
ISBN 978-7-04-018296-5

I. 计… II. ①范…②张…③姚… III. ①计算机应用-化学-英文
②分子结构-设计-英文 IV. O6-39 O561.1

中国版本图书馆 CIP 数据核字 (2008) 第 179705 号

策划编辑 李冰祥 责任编辑 柳丽丽 封面设计 王凌波 责任绘图 尹莉
版式设计 马敬茹 责任校对 胡晓琪 责任印制 陈伟光

出版发行	高等教育出版社	购书热线	010-58581118
社 址	北京市西城区德外大街 4 号	免费咨询	800-810-0598
邮政编码	100120	网 址	http://www.hep.edu.cn
总 机	010-58581000		http://www.hep.com.cn
		网上订购	http://www.landaco.com
经 销	蓝色畅想图书发行有限公司		http://www.landaco.com.cn
印 刷	涿州市星河印刷有限公司	畅想教育	http://www.widedu.com
开 本	787×1092 1/16	版 次	2009 年 3 月第 1 版
印 张	32	印 次	2009 年 3 月第 1 次印刷
字 数	540 000	定 价	68.00 元

本书如有缺页、倒页、脱页等质量问题,请到所购图书销售部门联系调换。

版权所有 侵权必究

物料号 18296-00

To my father and my mother

To my wife and my children

Botao Fan

Preface 1

The exponential development of computer technology since the late 1960s, not only provided chemists with extended computational capabilities but also gave rise to a new field: Chemoinformatics. Chemoinformatics which gathers all computerised treatments of chemical systems, at the confluence of several disciplines, knew how to take advantage of advances in such varied areas as: Computer Graphics, Theoretical Chemistry, Information Technologies, and rapidly became in the 1970s a well identified and worldwide recognized discipline with its own specific features.

High level quantum chemical calculations on the electronic structure of molecules are now possible in reasonable time. Time dependent evolution of complex systems (including solvent) can be characterized by Molecular Dynamics on a time scale allowing to grasp conformational changes, at least for medium-size molecules. Advances in Computer Graphics give varied, highly aesthetic, displays of molecular systems with interactive capabilities making the analyses more easily understandable and more user-friendly. Faced to the huge flow of data stored in databases or generated by combinatorial chemistry, Information Technologies provide efficient tools for exchanging, extracting and managing information.

To investigate chemical (or biological) properties or phenomena, numerous turnkey programs for molecular modeling and computational chemistry are now easily available, thanks particularly to the Internet. But to develop reliable and creative applications, it is essential that chemists have a good understanding of the models, their capabilities and physicochemical meaning, they must also be aware of their limitations and the approximation they are based on.

The authors, Prof. Fan B.T., Zhang R.S. and Yao J.H. joining their expertise in Physical Chemistry, Mathematics and Computer Science, present an excellent synthesis of complementary facets of molecular modeling, from theoretical background to applications.

The first two chapters of the book concern some relevant mathematical tools: fundamentals on linear algebra and differential equations and on another hand, minimization methods that is a key step in situations as diverse as searching for the preferred geometry of a molecular system or adjusting the parameterization of a model.

Chapters 3 to 5 introduce the basis of the quantum chemistry: *Ab Initio* Methods (including post-SCF treatments) are first presented. Density Functional Theory, and Semi-Empirical Approaches, at different levels of sophistication (from CNDO to PM3), are then developed and compared to the more traditional *Ab Initio* Methods.

Empirical Force Fields and Molecular Mechanics are then presented. Such methods are very rapid and now highly reliable and of great interest in conformational analysis of macromolecules (polypeptides, fragments of proteins, DNA...), which is a basic step in many applications.

The following chapters examine two powerful methods for the simulation of molecular systems: Molecular Dynamics and Monte Carlo Methods. They are widely used for the calculation of a lot of thermodynamic properties, such as free energies changes for varied processes (binding of a drug to a receptor, chemical reactions ...), or to take into account the influence of solvation.

The second part of the book is dedicated to applications of graph theory in Chemistry. Indeed Chemistry largely relies on the concept of structural formulae, that are not simple drawings but implicitly convey important structural information. Perception of these formulae as molecular graphs is fundamental for computer treatments and is at the very hearth of a number of approaches.

Extraction of substructures in a molecule, searching for a substructure common to several molecules in a set, recognition of ring systems are basic steps for efficient exploitation of databases. Detection of symmetry and equivalent sites in a structural formula is also a valuable tool which is the subject of chapters 10 to 13. This leads to a presentation of modern management systems for Chemical Information, a critical need on account of the huge flow of data. Formerly, particularly in drug design, the notion of similarity was essential: if two molecules are similar, it may be expected that they share the same properties. Now, as much as chemists more frequently work with large populations of molecules, rather than on limited collections, the notion of diversity takes increased interest. How comparable are two sets of structures? Does one of them more widely cover the structural space?

The third part of the book is dedicated to recent methods, relying on artificial intelligence that afford extended resources in many areas of chemical modeling: Artificial Neural Networks, Support Vector Machine and Genetic Algorithms. For example, in drug design, Neural Networks (NN) are able to extract from raw data (possibly with noise) the underlying (and sometimes complex) relationship between structures and an associated property, without need to define in advance a type of model. The various types of artificial neural networks are presented in details. Hopfield networks have been recently used for structural recognition. For classification tasks, Kohonen self organizing maps give a visual display making easier the perception of the results.

Genetic Algorithms mimicking the biological evolution process are essential for various applications. In the area of Quantitative Structure Activity Relationships (QSAR), the selection of relevant elements among the huge number of structural descriptors now available is a difficult and time-consuming problem, with often subjective decisions, whereas GAs offer a rational approach. Support vector ma-

chine (recently introduced) is designed for robust classifications or correlations with high predictive ability.

No doubt this book will provide undergraduate and graduate students with a sound training in molecular modeling, allowing them to master the actual state of the art and to easily adapt to the development of this rapidly evolving discipline.

Jean Pierre DOUCET
Professor. ITODYS. University Paris 7- Denis Diderot
September 2008, Paris

Preface 2

In 1997, the Chinese scientists working in France organized an activity to support the western development of China. We arrived at Lanzhou University. I had the opportunity to meet Professors HU Zhide, JIA Zhongjian, ZHENG Rongliang, LIU Mancang, and other teaching staffs of the Chemistry Department. The scientific exchange between us reached to a common point in research and scientific domains. Based on this same point of view, we signed a series of cooperation agreements, which include their requirement to open a summer class in Computer Chemistry for graduated students. This proposal obtained the support of the Ministry of Education of China. Professors DOUCET J.P. and PANAYE A. of University Paris 7-Denis Diderot expressed also their concerns and supports. Next year, 1998, the summer class was opened, this course of speciality was officially inserted in the list of teaching plan.

Encouraged by this successful beginning, Professor HU Zhide and Professor LIU Mancang suggested me to publish my course in order to solve the problem that the students lack teaching materials. After the discussion with Professor ZHANG Ruisheng, by considering the fact that the time is too short, we thought that we can not publish a book covering a large scope of this domain. Therefore we can only publish a book with limited contents.

Thanks to Lanzhou University, this teaching material was published in 1998. But because of the short time, it lacks a global plan. Moreover, the contents covered by this book are also limited, so we are not satisfactory for this version.

I'm now Professor of University Paris 7, and lead a research group, the laboratory of "Molecular Simulation and Molecular Information", in ITODYS institute. At the same time, I'm also the Director of Institute of Chemoinformatics of Lanzhou University. I have almost 20 years of experience in research and teaching of computer chemistry and published more than 120 papers and reviews in this area. My teaching experience covers almost all aspects of this area. Therefore, I hope to write a book which gathers the contents of our research and teaching work, in order to provide a reference to all researchers and graduated students worked in this domain.

Great thanks to High Education Press of China. Their heartily invitation gave me a good opportunity to realize my hopes. After receiving the invitation, I invited

two colleagues who work long time in Chemoinformatics, to work together for writing this book. One is Professor ZHANG Ruisheng of Lanzhou University, another is Associate Professor YAO Jianhua of Shanghai Institute of Organic Chemistry. The modern computer chemistry covers a large scope of contents. It includes the traditional computational chemistry, molecular graph theory and applications, chemoinformatics, molecular modeling, molecular design (including drug design), molecular simulation, QSAR/QSPR, and so on.

In this book, we cite and refer a lot of literatures reported by specialists. For example, in molecular modeling, personally I consider that the best book is "Molecular Modeling: Principles and Applications", written by A. Leach. In some related chapters of our book, we cited some contents of this reference book. About the mathematical principle of artificial neural network, we referred the book "Des réseaux de neurones", EYROLLES, 1990, written by E. Davalo and P. Naim. Massart *et al.* developed RBFNN (Massart is a great specialist in this area). The important statements about the theory and applications were reported in their published papers. These papers are the basis of RBFNN. In our book, some of their remarkable works have been cited and referred.

We cited also the works of our laboratory. These results are all the research works of my colleagues and PhD students, including Professor Doucet, Professor Panaye, Professor HU Zhide, Professor LIU Mancang, Dr. Barbault, Dr. Petitjean, Dr. Maldonado, Dr. CHEN Haifeng, Dr. YAO Xiaojun, Dr. XIA Hairong, Dr. LIU Huangxiang, Professor GAO Kun, and so on. I would like to mention specially Professor Doucet and Dr. Maldonado, because a very important part of their works has been cited in this book. I express here my sincere thanks to these colleagues and students. I thank also my family, my wife and children. Without their supports, encouragement and concerns, without a good environment of writing, I can not certainly finish the writing of this book.

This work was supported in part by National Natural Science Foundation of China through Grants 20572120.

I thank again the invitation and supports from Higher Education Press.

Botao FAN
August 2006, Paris

Contents

1	Numeric Analysis	1
1.1	Linear Algebraic Equations and Matrix	1
1.1.1	Square Matrix	2
1.1.2	Elementary Operations on Matrices	2
1.1.3	Inverse Matrices	6
1.2	Differential Equations	12
1.2.1	Cauchy Problem	12
1.2.2	Discrete One-Step Methods for Equations of First Order Principles	15
1.2.3	Discrete Methods with Multiple Steps for Differential Equations of Order 1	27
	Problems	35
2	Minimization	37
2.1	One-dimensional Minimization	39
2.1.1	Golden Section Search	39
2.1.2	Parabolic Interpolation: Brent's Method	41
2.2	Multidimensional Minimization	43
2.2.1	Downhill Simplex (Nelder et Mead, 1965)	43
2.2.2	Direction Set Method (Powell's Method)	44
2.2.3	Conjugate Directions	47
2.2.4	Powell's Quadratically Convergent Method: Find N Conjugate Directions	48
2.2.5	Gradient and the Steepest Descent Method	49
2.2.6	Newton-Raphson Method	53
	Problems	55
3	<i>Ab Initio</i>	56
3.1	Schrödinger Equation	56

3.2	Hartree-Fock Theory	58
3.3	Post-SCF Methods	63
3.3.1	Configuration Interaction	63
3.3.2	Moller-Plesset Perturbation Theory	64
3.4	Gaussian	67
3.4.1	Capabilities	67
3.4.2	Overview of Geometry Optimizations	68
3.4.3	Model Chemistries	69
3.4.4	Basis Sets	69
3.4.5	Limitations	72
3.5	Example of Applications	72
	Problem	78
4	Density Functional Theory and Applications	80
4.1	Theoretical Aspects	81
4.1.1	Hohenberg-Kohn Theorems	81
4.1.2	Kohn-Sham Theory	82
4.1.3	Exchange-Correlation Functionals	86
4.2	Comparison with Traditional <i>Ab Initio</i> Methods	91
4.3	Applications	92
4.3.1	Atoms	92
4.3.2	Clusters	93
4.3.3	Carbonyl Complexes	95
4.3.4	Other Applications	96
	Problems	100
5	Semi-Empirical Models	102
5.1	Extended Hückel Theory (EHT)	103
5.2	CNDO Method	105
5.3	INDO Method	108
5.4	MINDO Method	109
5.5	MNDO Method	110
5.6	AM1 and PM3 Methods	112
5.7	Example of Applications	112
5.7.1	Proposed Problem	112
5.7.2	Calculations	113
5.7.3	Conclusion	114
6	Molecular Mechanics	119
6.1	Stereochemistry	119
6.1.1	Constitution Isomers and Tautomers	119
6.1.2	Stereoisomers	120

6.2	Principle of Molecular Mechanics	122
6.3	Mathematical Expressions of Energy Terms	124
6.3.1	Bond Stretching Energy	125
6.3.2	Angle Bending Energy	126
6.3.3	Torsion Energy	127
6.3.4	Improper Torsions and Out-of-plane Bending Motions	129
6.3.5	Cross Terms: Class 1, 2 and 3 Force Fields	129
6.3.6	Energy of Non-binding Interactions	130
6.4	Force Fields	138
6.4.1	Parameterization Methods	139
6.4.2	SYBYL and TRIPOS Force Field	140
6.4.3	KOLLMAN Force Field	142
6.4.4	JUMNA and FLEX Force Field	142
6.5	Application Example	145
6.5.1	Problem to be Solved	145
	Problems	148
7	Molecular Dynamics	151
7.1	Molecular Dynamics with Simple Models	153
7.2	Molecular Dynamics with Continuous Potentials	154
7.3	Verlet Algorithm	154
7.4	Variations of Verlet Algorithm	155
7.5	Predictor-Corrector Integration Methods	157
7.6	Choosing the Time Step	158
7.7	Multiple Time Step Dynamics	159
7.8	Constraint Dynamics	160
7.9	Steps of Molecular Dynamics Simulation	164
7.10	Time-dependent Properties	165
7.11	Example of Applications	166
	Problems	171
8	Monte Carlo and Conformational Analysis	177
8.1	Monte Carlo	177
8.1.1	Calculations of Properties	179
8.1.2	Theory of Metropolis MC(MMC)	182
8.1.3	Implementation of MMC	184
8.1.4	Simulation of Molecules	185
8.2	Conformational Analysis	188
8.2.1	Conformational Search	189
8.2.2	Model-building Approaches	191
8.2.3	Random Search Methods	192

8.2.4	Distance Geometry	193
8.2.5	Other Approaches	196
	Problems	197
9	Molecular Graph and Presentation.....	198
9.1	Generality in Molecular Graph.....	198
9.2	Molecular Description.....	199
9.2.1	Computer Representation of Two-dimensional Chemical Structures	199
9.2.2	Connection Table	204
9.3	Sub-Structures and FREL	205
9.3.1	Extraction of FREL	205
9.3.2	Correlation of FREL with Properties.....	206
9.4	Canonical Numbering.....	209
9.5	Three-dimensional Representation of Molecular Structures	209
9.6	XML Representation.....	213
9.6.1	Principles of XML	214
9.6.2	XML for Structuring the Information	216
	Problem	219
10	Ring Perception	220
10.1	Graphs and Searching Methods	222
10.1.1	Graphs and Molecular Graphs	222
10.1.2	Depth-First Search.....	223
10.1.3	Breadth-First Search.....	225
10.2	Complete Ring Systems	226
10.2.1	Defined Terms	226
10.2.2	Number of Rings in SSSR.....	226
10.3	Zamora's Algorithm.....	229
10.4	Elimination Technique	230
	Problems.....	238
11	Constitutional Equivalence	239
11.1	Morgan Algorithm	240
11.2	Munk Algorithm	242
11.3	Rücker Algorithm	243
11.4	Varmuza Algorithm	245
11.5	Fan's Method	246
11.6	Application to Canonical Numbering.....	254
	Problem.....	255
12	Molecular Relative Symmetry.....	256
12.1	Relative Symmetry.....	257

12.2	Determination of Dissimilarity	259
12.2.1	Atom Focus	259
12.2.2	Bond Focus	259
12.3	Calculations	260
12.4	Examples	263
12.5	Application Examples: Empirical Rules	264
12.5.1	Empirical Rules for Gem Methyl Groups Adjacent to a sp^2 Carbon Site	265
12.5.2	Empirical Rules for Gem Methyl Groups Borne by a sp^3 Atom	266
	Problem	269
13	Isomorphism and 3D CSS Searches	270
13.1	Ullman's Algorithm	271
13.2	Crandell and Smith's Algorithm	273
13.3	Clique-Detection Algorithm	273
13.4	Lesk's Algorithm	275
13.5	Set-Reduction Algorithm	276
	Problems	278
14	Chemical Information Management and Exploration	280
14.1	Concept	280
14.2	Source of Chemical Information	282
14.2.1	Chemical Abstracts (CA) File	282
14.2.2	SCI Search	283
14.2.3	Chinese Scientific Citation Database (CSCD)	283
14.2.4	Science China Database (SCD)	283
14.2.5	SpecInfo	284
14.2.6	CrossFire	284
14.2.7	Cambridge Structural Database (CSD)	284
14.3	Management of Chemical Information	285
14.3.1	Principle	285
14.3.2	Representation of Chemical Structure	286
14.3.3	Chemical Structure Searching	288
14.3.4	Examples	288
14.4	Data Mining and Spectral Simulation	290
14.4.1	Data Mining	290
14.4.2	Application: IR/Raman Simulation	291
14.4.3	Application: Mass Spectra Simulation	298
	Problems	313

15	Molecular Similarity and Diversity	314
15.1	Approaches on Molecular Similarity and Diversity	316
15.1.1	Components of the Molecular Similarity	316
15.1.2	The Similarity Paradox	319
15.1.3	Similarity/Diversity Approaches in Chemoinformatics	322
15.2	Selection, Classification and Validation Methods	337
15.2.1	Classification Methods	338
15.2.2	Selection Methods	341
15.2.3	Validation Methods	342
15.3	Comparative Analysis of Approaches/Descriptors-Models	
	Comparison	344
15.3.1	Comparison of Similarity Coefficients	344
15.3.2	Choice of Descriptors	345
15.3.3	Comparison and Selection of Descriptors	345
15.4	Applications in Chemo- and Bio- Informatics	348
15.4.1	Combinatorial Chemistry, Chemical Libraries	348
15.4.2	Drug Design and Medicinal Chemistry	350
15.4.3	Synthesis Chemistry, Organic Chemistry and Catalyst	352
15.4.4	Bioinformatics	352
	Problems	353
16	Artificial Neural Networks: Biological Fundamental and Modeling	354
16.1	Introduction	354
16.2	Biological Fundamental	355
16.3	Modeling	356
16.4	Structure of Connections	359
	Problems	363
17	Layered Network	364
17.1	Mono-Layered Associative Network	364
17.1.1	Perceptron	364
17.1.2	Linear Separation	365
17.1.3	Limitations	366
17.1.4	Widrow-Hoff Rule	366
17.2	Back-Propagation Learning Algorithm	368
17.2.1	Example: Prediction of ^{13}C -NMR Shifts for Methyl Substituted Cyclohexanes	373
17.3	Radial Basis Function Networks	375
17.3.1	Theory	375

17.3.2	Training Algorithms	378
17.3.3	Algorithms for the Radial Basis Functions Selection	380
17.3.4	Design of Training Data Set for the Calibration Problem	382
17.3.5	Example of Applications	383
	Problems	412
18	Hopfield Network	413
18.1	Modeling	413
18.1.1	Network Completely Connected	413
18.1.2	Leaning Process in a Hopfield Network	414
18.1.3	Stability of States	415
18.2	Application to Optimization Procedures	417
18.2.1	Energy of Hopfield Network	417
18.2.2	Traveling Salesman Problem (TSP)	418
18.2.3	3D Structural Recognition	419
18.3	Boltzmann Machine	422
	Problems	423
19	Kohonen Network	425
19.1	Structure of SOM	425
19.2	The Learning Process	426
19.2.1	Initialization	427
19.2.2	Training	427
19.2.3	Variants	429
19.3	Examples and Applications	430
19.3.1	Mapping the 3D Sphere Surface onto a SOM Plane	430
19.3.2	Application to Visual Classification of Gas Chromatographic Liquid Phase	432
	Problem	438
20	Genetic Algorithms and Their Applications in Chemistry	441
20.1	Optimization and Techniques	442
20.2	Simple Genetic Algorithm	444
20.2.1	Population and Coding	444
20.2.2	Fitness Function	445
20.2.3	Operators	446
20.3	Mathematical Fundamental	449
20.4	Application of SGA to QSAR Problem	453
20.5	Other Application of GA to Chemistry Problems	455
20.5.1	Conformational Searching	455
20.5.2	Other Applications	457

Problems	458
21 Support Vector Machine (SVM)	460
21.1 SVM in Classification	461
21.1.1 Basic Principle	462
21.1.2 Some Applications	468
21.2 Regression SVM	476
21.2.1 Basic Principle	477
21.2.2 Further Developments	479
21.3 Applications	480
Problems	490