On Using Automated Algorithms to Parameterise Molecules for Molecular Dynamics Simulations and Investigating Suitable Ensembles for the Simulation of Naphthalimide Monolayers

Ivan Welsh

MASSEY UNIVERSITY
Te Kunenga Ki Purehuroa

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

Institute of Natural and Mathematical Sciences

A thesis
submitted to Massey University in Albany, Auckland
in fulfilment of the requirements for the degree of
Doctor of Philosophy
in Chemistry.

Massey University Auckland
2017

Abstract

Molecular dynamics simulations provide a means to investigate the spatial and temporal evolution of systems of molecules at atomic resolution. Force fields are used to describe the interactions between atoms contained within the system. A number of such force fields have been developed over the years, with a focus on force fields for use in simulations of biochemical systems, in particular, protein systems. This thesis is primarily focused on extending the range of systems that can be simulated through providing means for automated generation of force field parameters for large novel molecules.

One component of existing force fields that is generally poorly parameterised are the dihedral terms. In combination with the non-bonded terms, the dihedral terms are used to describe the rotational energy profile about bonds, and have a large influence on the conformational properties of a simulated system. A new method for the determination of dihedral parameters is developed, utilising high level quantum mechanical calculations. With the use of local elevation molecular dynamics simulations, this method is applied to the case of protein backbone dihedrals within the GROMOS force field.

When one desires to simulate the interaction of a novel molecule with some biochemical system, the novel molecule must be parameterised in a manner that is compatible with the force field used to describe the biochemical system. However, doing so is a slow, tedious, and error prone process, especially when the novel molecule is large. To combat this, a new algorithm, known as CherryPicker, was developed. CherryPicker is a graph based algorithm which enables rapid parameterisation of large molecules through fragment comparison with a library of previously parameterised small molecules. The algorithm design is discussed and tested on a few simple test cases in part II.

Part III steps away from the parameterisation focus of this thesis and looks at the simulation of naphthalimide monolayers. Naphthalimides have applications in sensing environments as they have absorption and fluorescence emission spectra lying within the UV and visible regions of light. With a long chain alkane substituted at the N-imide site, they become amphiphilic and can form monolayers on the surface of water, and can be transferred to a solid substrate when at a desired compression level. Molecular dynamics simulations can be used to provide insight into the formation of compressed monolayer phase. Here, the effect of different ensembles, namely NVT, NPT, and NγT are investigated for use in simulating a naphthalimide monolayer.

Acknowledgements

Lets keep this short. First person to thank is Dr Jane Allison, for hiring me as a PhD Student, and paying me, and basically letting me choose what work to do, even though it involved lots of things outside of her area of expertise. Also thanks to Professor Alan Mark and his group at the University of Queensland, particularly Martin and Bertrand, for hosting me a couple of times. To Dr yomcat (little y), thanks for dealing with my stoopid maths questions as they arose. And finally, thanks to the volunteer proofreaders who gave up their time to read my brain dumps. So cheers to Dr Thomas Collier, Bruce Welsh, and Nina Leeb. Nina, I look forward to reciprocating when its your time. Oh and big thanks to my lovely mother who decided to send me fantastic lunches three days a week. Because physics. Finally, because I forgot them in my examination copy, thanks to my group members for all the useful stuff anf things over the years, particularly Ashar.

Preface

With a preceding introductory chapter, this thesis is divided into three distinct parts. Part I focuses on the SpinningTop program, which is a program developed for determining dihedral parameters. A brief background to the reasons that such a method is required is given in chapter 2. Chapter 3 outlines the theory and implementation of the fitting method, and investigates some of the considerations that need to be made. Chapter 4 details the methods used to translate the developed fitting method to the case of protein backbone dihedral terms within the GROMOS force field, and chapter 5 discusses the results obtained in this proof of principle work.

Part II of the thesis focusses on the CherryPicker algorithm, which is a new algorithm developed to enable easy parameterisation of large biochemical molecules compatible with the GROMOS force field. As the algorithm is based on the concept of molecular fragmentation, a brief introduction to the state of computational molecular fragmentation is given in chapter 6. The design and mathematical background of the algorithm is presented in chapter 7, before a small amount of proof-of-concept testing is undertaken in chapter 8. As part of the CherryPicker algorithm development, a novel means to automatically determine bond order and formal charges of molecules was developed. This is presented in chapter 9.

Finally, part III presents work undertaken in the determination of suitable ensembles for the simulation of naphthalimide monolayers.

A large amount of code was developed as part of the work for this thesis. This code is available on request to j.allison@massey.ac.nz. The code will be provided as is, with no documentation on system requirements, installation, or usage.

Contents

Preface v

List of Figures xv

List of Tables xvii

1. Introduction 1
   - 1.1. Molecular Dynamics 1
     - 1.1.1. Integration 2
     - 1.1.2. Initialisation 2
   - 1.2. Force Fields 6
     - 1.2.1. Bonded Terms 8
     - 1.2.2. Non-bonded Terms 11
   - 1.3. Other Force Field Types 16
   - 1.4. Free Energy 18
   - 1.5. Local Elevation 19
   - Bibliography 20

I. SpinningTop 29

2. Introduction 31

3. Theory and Method Development 33
   - 3.1. Derivation of Fitting Method 33
     - 3.1.1. Fitting Symmetric Dihedrals 33
     - 3.1.2. Fitting with Phase Shift 35
   - 3.2. Robust Regression 37
   - 3.3. Electronic Effects of Distance of Substituents from Dihedral 39
   - 3.4. Sampling Density Requirements 42

4. Method 45
- 4.1. Amino Acid Choices 45
- 4.2. Dihedral Parameter Fitting 48
  - 4.2.1. Quantum Chemical Calculations 48
  - 4.2.2. Molecular Dynamics Simulations 49
  - 4.2.3. Parameter Fitting 50
- 4.3. Experimental Comparisons 50
  - 4.3.1. Relative Energy Calculation 51
  - 4.3.2. Secondary Structure 51

5. Results and Discussion 53
- 5.1. Fitting Outcomes 53
- 5.2. General Comments 64
- 5.3. Conclusions 64

Bibliography 65

II. CherryPicker 69

6. Introduction 71
- 6.1. Automated Molecular Parametrisation 71
  - 6.1.1. Rule Based Approaches 72
  - 6.1.2. Quantum Mechanics Based Approaches 72
  - 6.1.3. Novel Force Field Generation Approaches 73
- 6.2. Molecular Fragmentation 74

7. Algorithmic Design and Theory 77
- 7.1. Mathematical Concepts 77
  - 7.1.1. Set Definitions 77
  - 7.1.2. Graph Theoretic Definitions 79
- 7.2. Condensed Molecular Graph 83
  - 7.2.1. Vertex Colours 84
  - 7.2.2. Edge Colours 85
- 7.3. Stereochemical Determination 86
  - 7.3.1. CIP String Generation 86
  - 7.3.2. R/S Chirality Determination 86
  - 7.3.3. E/Z Geometric Isomerisation Determination 87

7.4. Athenaeum 87
7.4.1. Fragment definition 87
7.4.2. Dihedral Fragments 92
7.5. Subgraph Isomorphism Testing 93
7.6. Parameterisation 94
7.6.1. Condensed Atoms 94
7.6.2. Symmetry 95
7.6.3. Non-bonded Terms 95
7.6.4. Bond and Angle Terms 101
7.6.5. Proper and Improper Dihedral Terms 102
7.6.6. Unmapped regions 103

8. Testing and Discussion 105
8.1. Algorithm Optimisation 105
8.1.1. Charge Group Partitioning 105
8.1.2. Fragment Generation 110
8.1.3. Fragment Size 113
8.1.4. Overlap 113
8.2. Parameterisation Tests 116
8.2.1. Structural Minimisation 118
8.2.2. Dynamic Stability 119
8.2.3. Nuclear Magnetic Resonance 121
8.3. Conclusion 128

9. Bond Order and Formal Charge Assignment 129
9.1. Introduction 129
9.2. General Problem Information 131
9.2.1. Optimisation function 131
9.2.2. Electron Count 132
9.2.3. Electron Positioning 132
9.3. Energy Calculations 133
9.3.1. Formal Charge Energies 133
9.3.2. Bond Order Energies 137
9.3.3. Charged Bonds 140
9.3.4. Lookup Tables 140

9.4. Optimisation Methods 141
9.4.1. Local Optimisation 141
9.4.2. Genetic Algorithm 142
9.4.3. A* 145
9.4.4. Fixed Parameter Tractable (FPT) 147
9.4.5. Evaluation 152
9.5. Conclusion 157

Bibliography 158

III. Monolayers 167

10. Introduction 169
10.1. Naphthalimides 169
10.2. Monolayers 170
10.2.1. Monolayer Simulation 171

11. Methods 173
11.1. Parameterisation 173
11.2. Computational Details 173
11.2.1. System Construction 173
11.2.2. Simulation Conditions 174
11.2.3. NPT Simulations 174
11.2.4. NγT Simulations 174
11.2.5. NVT Simulations 174
11.3. Surface Pressure Calculation 175

12. Results and Discussion 177
12.1. Monolayer Structural Properties 178
12.2. Conclusion 182

Bibliography 183

13. Summary and Future Endeavours 187
13.1. SpinningTop 187
13.2. CherryPicker 188
13.3. Monolayers 189

IV. Appendices 191

A. Dihedral Energy Profiles 193

B. Sampling Density RMSD Values with Different Sample Sets 199

C. PDB Ramachandran Plots 205

D. Raw Energy Surfaces 213

E. Naphthalimide Parameters 219

F. ATB Molecules in SRC9064 227

G. Bond Order Assignment Validation Molecules 251

H. Electron Position Probabilities 281

Glossary 283

List of Figures

1.1. Periodic boundary conditions as applied to a box of water 4
1.2. Diagrammatic representation of force field terms 7
1.3. Affect of symmetry on calculated partial atomic charges of benzene 15

2.1. Schematic of the Φ and Ψ amino acid backbone dihedrals 31

3.1. Diagramatic representation of addition of vectors and their projection onto the x-axis 36
3.2. Using robust regression to fit to data with outliers 39
3.3. 3-ethyl hexane 40
3.4. Fits to substituted dihedral rotational energy profiles compared with the unsubstituted energy profile 41

4.1. Structures of the capped amino acids used for dihedral parametrisation 48

5.1. Fitted terms and Φ/Ψ surface plots for glycine dipeptide 54
5.2. Fitted terms and Φ/Ψ surface plots for alanine dipeptide 55
5.3. Fitted terms and Φ/Ψ surface plots for valine dipeptide 57
5.4. Fitted terms and Φ/Ψ surface plots for serine dipeptide 58
5.5. Fitted terms and Φ/Ψ surface plots for cysteine dipeptide 59
5.6. Fitted terms and Φ/Ψ surface plots for glutamine dipeptide 60
5.7. Fitted terms and Φ/Ψ surface plots for phenylalanine dipeptide 61
5.8. Fitted terms and Φ/Ψ surface plots for aspartic acid dipeptide 62
5.9. Fitted terms and Ψ profile plots for proline dipeptide 63

6.1. Components of a phospholipid 75

7.1. Schematic representation of the CherryPicker algorithm. 78
7.2. Venn diagrams of set operations 79
7.3. An example graph 80
7.4. A graph G with subgraphs G' and G'' 80

7.5. A path $P = P^6$ in G. 81
7.6. A cycle 81
7.7. A three component graph 82
7.8. A tree with root r. 82
7.9. Bit string representation of the vertex colour 32-bit integer 84
7.10. Bit string representation of the edge colour 8-bit integer 85
7.11. Fragment of a graph with overlap regions 91
7.12. Example of the DAG produced from a fragment set 92
7.13. A line graph 101

8.1. Structures of the CherryPicker test molecules 106
8.2. Charge group size distributions with various w values 108
8.3. Charge group charge distributions with various w values 109
8.4. Charge group diameter distributions with various w values 111
8.5. Edge terminated fragment of a graph with overlap regions 112
8.6. Distributions of mapped bond parameter counts 114
8.7. Distributions of mapped angle parameter counts 115
8.8. Charge distributions of simple atomic fragment and overlap combinations . . 117
8.9. Molecule XVI with soft bonds marked in red 121
8.10. Experimental ^{1}H NMR spectra 123
8.11. ^{1}H NMR spectra of molecule I 124
8.12. ^{1}H NMR spectra of molecule XXII 124
8.13. ^{1}H NMR spectra of molecule IX 125
8.14. ^{1}H NMR spectra of molecule VII 125
8.15. ^{1}H NMR spectra of molecule VIII 126
8.16. ^{1}H NMR spectra of molecule XVI 126
8.17. ^{1}H NMR spectra of molecule X 127
8.18. Time series of amide dihedral during simulation 127

9.1. A graph with a tree-decomposotion and a nice tree-decomposition 149

10.1. 1,8-naphthalimide structure 169
10.2. Idealised pressure-area isotherm 170
10.3. Structure of the naphthalimide molecule investigated here. 172

12.1. Pressure–area isotherms plots the naphthalimide monolayer simulated in three different ensembles 177

12.2. Simulation snapshots 179
12.3. Progression of deuterium order parameters upon monolayer compression . . . 180
12.4. Radial distribution function of carbon nine in the alkyl chain upon compression 181

List of Tables

3.1. Sampling density RMSD results 43

4.1. Amino acid counts 46
4.2. RMSD values between normalised Ramachandran plots for all twenty natural amino acids. 47

5.1. Reference GROMOS 54A7 backbone dihedral parameters for all amino acids 54

8.1. Successfully parameterised molecules 118
8.2. RMSD of minimised test molecules 119
8.3. Overall molecular bond length and angle percentage deviations 121

9.1. Relative atomic energies for elements with formal charges 134
9.2. Negative bond dissociation energies 138
9.3. Charged bond dissociation energies 141
9.4. Efficency and accuracy of algorithms 154
9.5. Comparison to reference assignments 156

List of Abbreviations

AA all atom.

ATB Automated Topology Builder.

BSSE basis set superposition error.

CIP Cahn–Ingold–Prelog priority rules.

CSD Cambridge Structural Database.

DAG directed acyclic graph.

FPT fixed parameter tractable.

GAFF Generalised AMBER force field.

NBO Natural Bond Orbital.

PDB Protein Data Bank.

QM quantum mechanical.

QMDFF Quantum Mechanically Derived Force Field.

RMSD root-mean-square deviation.

UA united atom.

UFF Universal Force Field.