Boundary value approaches to molecular dynamics simulation

by

Peter Thomas Vedell

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Co-Majors: Bioinformatics and Computational Biology; Applied Mathematics

Program of Study Committee:
Zhijun Wu, Co-major Professor
Robert Jernigan, Co-major Professor
Kai-Ming Ho
Howard Levine
Michael Smiley
Ron Elber

Iowa State University
Ames, Iowa
2007

Copyright © Peter Thomas Vedell, 2007. All rights reserved.
# TABLE OF CONTENTS

LIST OF ACRONYMS ............................................................................................................................................ xi
LIST OF TABLES ................................................................................................................................................... xiii
LIST OF FIGURES ................................................................................................................................................ xiv
LABELING OF MATHEMATICAL EXPRESSIONS AND USAGE OF SYMBOLS ........................................ xv
LIST OF REDUNDANCIES .................................................................................................................................. xvi
ABSTRACT ........................................................................................................................................................... xvii

1 OVERVIEW ........................................................................................................................................................ 1

1.1 Introduction ..................................................................................................................................................... 1

1.2 Preliminaries .................................................................................................................................................. 3

1.3 Overview of doctoral research ..................................................................................................................... 10

1.4 References .................................................................................................................................................... 16

2 BOUNDARY VALUE BIOMOLECULAR DYNAMICS SIMULATION AND NUMERICAL METHODS .......... 17

2.1 Preliminaries I ................................................................................................................................................. 17

2.1.1 Potential energy functions .......................................................................................................................... 17

2.1.2 Metrics ....................................................................................................................................................... 18

2.1.3 Local minima .......................................................................................................................................... 18

2.1.4 All-atom molecular dynamics simulation (AA-MDS) ........................................................................... 19

2.1.5 Initial value AA-MDS (IV- AA-MDS) ....................................................................................................... 19

2.1.6 Boundary value AA-MDS (BV- AA-MDS) .............................................................................................. 20

2.1.7 Generalized coordinates; relating potential energy functions and AA-MDS ...................... 20
2.2 Ideas, Methods, and Analysis I ................................................................. 23

2.2.1 Specification of boundary conditions for AA-MDS ........................................... 23

2.2.2 Hamiltonians, equilibrium points, stability, and wells of a potential energy surface... 25

2.3 Preliminaries II ......................................................................................... 30

2.3.1 Multiple shooting ................................................................................... 30

2.3.1.1 Example 1: Transformation from 2nd-order to 1st-order BVP ....... 31

2.3.1.2 Example 2: Determining the Jacobian .............................................. 35

2.3.1.3 Example 3: Application to Mueller potential ...................................... 36

2.3.2 Single shooting .................................................................................... 37

2.3.3 Finite difference methods ........................................................................ 37

2.3.4 Global convergence approaches ........................................................... 40

2.3.5 Identifying a numerical solution ............................................................. 42

2.3.6 Globally convergent modifications of Newton’s method ......................... 42

2.3.6.1 Trust-region model with dogleg step .............................................. 43

2.3.6.2 Damped Newton model with the natural criterion function ............. 45

2.3.7 Stochastic difference equation methods .................................................. 46

2.3.8 Background information for biomolecular simulation ............................. 48

2.3.8.1 Classical mechanics and quantum mechanics ...................................... 48

2.3.8.2 Characterization of protein conformations ....................................... 49

2.3.8.3 System of units .............................................................................. 51

2.3.9 Computational limitations of AA-MDS and implications ....................... 51

2.3.10 Distance matrix interpolation ............................................................... 53

2.3.10.1 Example 4: Linear DMI for transformation of a triangle .............. 54

2.3.10.2 Example 5: Nonlinear DMI for transformation of a triangle ......... 54

2.4 Ideas, Methods, and Analysis II ............................................................... 54
| 2.4.1 | Why AA-MDS? ........................................................................................................... | 54 |
| 2.4.2 | Why BV-AA-MDS? .................................................................................................... | 55 |
| 2.4.3 | Why multiple shooting? .......................................................................................... | 63 |
| 2.4.3.1 | Challenging aspects of single shooting ............................................................... | 63 |
| 2.4.3.2 | Challenging aspects of finite difference methods ................................................. | 68 |
| 2.4.3.3 | Example 6 : Conditioning of scalar linear 2nd-order BVP and IVP .......................... | 69 |
| 2.4.3.4 | Example 7 : Shadowing lemma and conditioning of BVP ................................. | 70 |
| 2.4.3.5 | Example 8 : Number of variables for application of finite differences ............... | 72 |
| 2.4.3.6 | Challenging aspects of multiple shooting methods .............................................. | 72 |
| 2.4.3.7 | Example 9 : Conditioning, variable count for application of multiple shooting ..... | 73 |
| 2.4.4 | Relative computational cost of BV-AA-MDS and IV-AA-MDS .............................. | 73 |
| 2.4.5 | Comparison with Stochastic Difference Equation approach .................................. | 75 |
| 2.5  | Summary ..................................................................................................................... | 78 |
| 2.6  | Appendix: Order of accuracy of selected numerical methods for solution of 2nd-order Hamiltonian systems ........................................................................ | 80 |
| 2.6.1 | Phase flow, discrete evolution, consistency, and convergence ............................. | 80 |
| 2.6.2 | Central Verlet algorithm : position ..................................................................... | 81 |
| 2.6.3 | Central Verlet algorithm : velocity .................................................................... | 82 |
| 2.6.4 | Non-central Verlet algorithm : position ............................................................... | 82 |
| 2.6.5 | Velocity Verlet algorithm : position ................................................................... | 83 |
| 2.6.6 | Velocity Verlet algorithm : velocity ................................................................... | 83 |
A MULTIPLE SHOOTING APPROACH FOR ALL-ATOM SIMULATION OF TRANSITIONS BETWEEN TWO MOLECULAR CONFORMATIONS SPECIFIED BY SETS OF INTERNAL COORDINATES

3.1 Abstract

3.2 Introduction

3.3 Preliminaries

3.3.1 Current approaches to the study of conformational transitions of proteins

3.3.1.1 Approaches to generate ordered, accurate trajectories

3.3.1.2 Approaches to generate detailed, ordered, approximate trajectories

3.3.1.3 Approaches to generate coarse-grained, ordered, approximate trajectories

3.3.2 All-atom molecular dynamics simulation (AA-MDS)

3.3.3 Initial value AA-MDS (IV-AA-MDS)

3.3.4 Boundary value AA-MDS (BV- AA-MDS)

3.3.5 Multiple shooting (MS)

3.3.6 Single shooting (SS)

3.3.7 Numerical solution of initial value problems (IVP’s)

3.3.8 Global convergence algorithms
3.4 Ideas, methods, and analysis

3.4.1 Specification of boundary conditions for AA-MDS

3.4.2 N-acetyl-N′-methylalaninamide - a blocked alanine dipeptide

3.4.3 Toward a parameter selection strategy

3.4.3.1 Assumptions about simulations with same initial velocity direction

3.4.3.2 A strategic approach for finding an approximate BV-AA-MDS solution and estimating the time interval for a conformational transition

3.4.3.3 Example 10: Utility of strategic for finding an approximate BV-AA-MDS solution

3.4.3.4 Quantitative criteria for identification of approximate BV-AA-MDS solution

3.4.3.5 Finding initial parameter vector candidates by directly using an approximate BV-AA-MDS solution

3.4.3.6 Finding initial parameter vector candidates by directly using a BV-AA-prx trajectory

3.4.3.7 Finding an augmented initial parameter set using a hidden time BV-AA-prx trajectory

3.4.3.8 Example 11: Finding an augmented initial parameter set using a hidden time BV-AA-prx trajectory

3.4.3.9 Restarting

3.4.4 A MS BV-AA-MDS strategy for min→min transitions

3.4.5 Applying the MS strategy to conformational transitions of N-acetyl-N′-methylalaninamide

3.4.6 Additional analysis of N-acetyl-N′-methylalaninamide study and further discussion
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.3.4</td>
<td>Single shooting (SS)</td>
<td>178</td>
</tr>
<tr>
<td>4.3.5</td>
<td>All-atom normal mode analysis</td>
<td>180</td>
</tr>
<tr>
<td>4.3.6</td>
<td>Distance measures and distance geometry</td>
<td>182</td>
</tr>
<tr>
<td>4.4</td>
<td>Ideas, methods, and analysis</td>
<td>183</td>
</tr>
<tr>
<td>4.4.1</td>
<td>Motivation for boundary conditions based on lower and upper bounds for interatomic distances</td>
<td>183</td>
</tr>
<tr>
<td>4.4.2</td>
<td>Defining boundary conditions based on lower and upper bounds for interatomic distances</td>
<td>184</td>
</tr>
<tr>
<td>4.4.3</td>
<td>Defining a boundary condition for an upper bound on total energy</td>
<td>185</td>
</tr>
<tr>
<td>4.4.4</td>
<td>Fixing a subset of initial parameters to satisfy initial boundary conditions of separable BVP</td>
<td>186</td>
</tr>
<tr>
<td>4.4.5</td>
<td>Parameterization using normal mode directions</td>
<td>187</td>
</tr>
<tr>
<td>4.4.6</td>
<td>Parameterization using a subset of normal mode directions</td>
<td>188</td>
</tr>
<tr>
<td>4.4.7</td>
<td>Parameter update methods</td>
<td>189</td>
</tr>
<tr>
<td>4.4.8</td>
<td>A SS BV-AA-MDS strategy for \textit{well} \rightarrow \textit{well} transitions</td>
<td>191</td>
</tr>
<tr>
<td>4.4.9</td>
<td>Applying the SS strategy to study of conformational transitions of</td>
<td>195</td>
</tr>
<tr>
<td></td>
<td>\textit{N-acetyl-N'}-methylalaninamide</td>
<td></td>
</tr>
<tr>
<td>4.4.10</td>
<td>Additional analysis of \textit{N acetyl N'}-methylalaninamide study and further discussion</td>
<td>198</td>
</tr>
<tr>
<td>4.5</td>
<td>Summary</td>
<td>199</td>
</tr>
<tr>
<td>4.6</td>
<td>References</td>
<td>200</td>
</tr>
<tr>
<td></td>
<td><strong>Figures and Tables</strong></td>
<td>204</td>
</tr>
<tr>
<td>5</td>
<td>DISTANCE MATRIX INTERPOLATION METHODS FOR BOUNDARY VALUE APPROACHES TO BIOMOLECULAR DYNAMICS SIMULATION</td>
<td>220</td>
</tr>
<tr>
<td>5.1</td>
<td>Abstract</td>
<td>220</td>
</tr>
</tbody>
</table>
5.2 Introduction .................................................................................................................. 221

5.3 Preliminaries: All-atom distance matrix interpolation using an elastic network model with linear polynomial interpolation for target distance matrices (linear AA-DMI-ENM) .................................................................................................................. 224

5.3.1 Defining an optimization problem for linear distance matrix interpolation ........224
5.3.2 Simplifying the optimization problem by a quadratic approximation to the objective function ................................................................................................................. 226
5.3.3 Generation of AA-DMI-ENM trajectories in three-dimensional space .......... 227
5.3.4 Strategies for selection of subset of interatomic distances .............................. 228
5.3.5 Strategies for assignment of interpolation weight, or spring constants .......... 229

5.4 Ideas, methods and analysis ..................................................................................... 230
5.4.1 Identifying and evaluating assumptions of linear DMI-ENM ......................... 230
5.4.2 Application of linear AA-DMI-ENM to conformational transitions of

N-acetyl-N'-methylalaninamide .......................................................... 231

5.4.2.1 Feasibility of the linear AA-DMI-ENM trajectories .................... 232
5.4.2.2 Assessing the linearity assumption ........................................ 232
5.4.2.3 Assessing the quadratic approximation assumption .................. 233
5.4.2.4 Additional considerations .................................................... 234
5.4.3 An alternative approach to DMI: All-atom distance matrix interpolation using the

STRAIN() objective function ................................................................. 235
5.4.4 Interpolation by cubic spline functions for target distance matrices ........... 238
5.4.5 Refining trajectories using local constrained energy minimization (LCEM) ....... 240
5.4.6 Application of spline interpolation and AA-DMI-DG to conformational transitions of

N-acetyl-N'-methylalaninamide ................................................................. 241
5.4.6.1 Using the STRAIN() objective function with consistent target
distance matrices ................................................................. 241
5.4.6.2 Local constrained energy minimization (LCEM) .................... 242
5.4.6.3 Spline interpolation for target distance matrices ..................... 242
5.4.7 Additional analysis and further discussion ..................................... 243

5.5 Summary ................................................................................................. 247

5.6 References ............................................................................................... 247

Figures and Tables .......................................................................................... 250

6 SUMMARY AND FUTURE DIRECTIONS ............................................................. 264

6.1 Summary .................................................................................................... 264

6.2 Future directions in applications ............................................................... 267

6.3 Future directions in methods and analysis ................................................... 268

ACKNOWLEDGEMENTS .................................................................................... 270
# LIST OF ACRONYMS

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Term or Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA-DMI</td>
<td>All-Atom Distance Matrix Interpolation</td>
</tr>
<tr>
<td>AA-DMI-DG</td>
<td>All-Atom Distance Matrix Interpolation</td>
</tr>
<tr>
<td>AA-DMI-DG-LCEM</td>
<td>All-Atom Distance Matrix Interpolation using Distance Geometry approach with trajectory refinement by Local Constrained Energy Minimization</td>
</tr>
<tr>
<td>AA-DMI-ENM</td>
<td>All-Atom Distance Matrix Interpolation</td>
</tr>
<tr>
<td>AA-MDS</td>
<td>All-Atom Molecular Dynamics Simulation</td>
</tr>
<tr>
<td>AMBER</td>
<td>Assisted Model Building And Energy Refinement (a family of potential energy functions and force fields)</td>
</tr>
<tr>
<td>BV-AA-MDS</td>
<td>Boundary Value All-Atom Molecular Dynamics Simulation, or Boundary Value approach to All-Atom Molecular Dynamics Simulation</td>
</tr>
<tr>
<td>BVP</td>
<td>Boundary Value Problem</td>
</tr>
<tr>
<td>CPMD</td>
<td>Coupled Parallel Molecular Dynamics simulation</td>
</tr>
<tr>
<td>DDD</td>
<td>Dynamic Distance Distribution</td>
</tr>
<tr>
<td>DG</td>
<td>Distance Geometry</td>
</tr>
<tr>
<td>DMI</td>
<td>Distance Matrix Interpolation. This acronym and acronyms beginning with DMI may be equivalent to the same acronym AA added as a prefix when the all-atom aspect is clear from context.</td>
</tr>
<tr>
<td>DMI-DG</td>
<td>Distance Matrix Interpolation using Distance Geometry approach</td>
</tr>
<tr>
<td>DMI-DG-LCEM</td>
<td>Distance Matrix Interpolation using Distance Geometry approach with trajectory refinement by Local Constrained Energy Minimization</td>
</tr>
<tr>
<td>DMI-ENM</td>
<td>Distance Matrix Interpolation using Elastic Network Modeling approach</td>
</tr>
<tr>
<td>ENM</td>
<td>Elastic Network Model</td>
</tr>
<tr>
<td>IV-AA-MDS</td>
<td>Initial Value All-Atom Molecular Dynamics Simulation, or Initial Value approach to All-Atom Molecular Dynamics Simulation</td>
</tr>
<tr>
<td>IVP</td>
<td>Initial Value Problem</td>
</tr>
<tr>
<td>LCEM</td>
<td>Local Constrained Energy Minimization</td>
</tr>
<tr>
<td>MATLAB</td>
<td>MATrix LAboratory (a software package)</td>
</tr>
<tr>
<td>MDS</td>
<td>Molecular Dynamics Simulation</td>
</tr>
<tr>
<td>$\min \rightarrow \min$</td>
<td>A transition between local minima of a potential energy surface. If two local minima are named $A$ and $B$, then the notation $A \rightarrow B$ refers to a transition from local minimum $A$ to local minimum $B$.</td>
</tr>
<tr>
<td>MS</td>
<td>Multiple Shooting</td>
</tr>
<tr>
<td>NLE</td>
<td>NonLinear Equation</td>
</tr>
<tr>
<td>NMA</td>
<td>Normal Mode Analysis</td>
</tr>
<tr>
<td>ODE</td>
<td>Ordinary Differential Equation</td>
</tr>
<tr>
<td>REMD</td>
<td>Replica Exchange Molecular Dynamics simulation</td>
</tr>
<tr>
<td>SDE</td>
<td>Stochastic Difference Equation</td>
</tr>
<tr>
<td>type of trajectory: approximate BV-AA-MDS trajectory</td>
<td>A trajectory that satisfies discretized Newtonian equations of motion and approximately satisfies boundary conditions of a BVP.</td>
</tr>
<tr>
<td>type of trajectory: BV-AA-MDS solution, or BV-AA-MDS solution trajectory</td>
<td>A trajectory that satisfies conditions of a BVP. So, it satisfies boundary conditions and also satisfies discretized Newtonian equations of motion.</td>
</tr>
<tr>
<td>Acronym</td>
<td>Term or Explanation</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>type of trajectory: BV-AA-<em>prx</em> trajectory</td>
<td>A trajectory generated by an arbitrary approximate method that satisfies boundary conditions of a BVP does not necessarily satisfy Newtonian equations of motion. For example, a BV-AA-DMI trajectory is a type of BV-AA-<em>prx</em> trajectory. In Chapter 5, we refer to a BV-AA-DMI trajectory as an AA-DMI trajectory since a DMI trajectory implicitly satisfies specified boundary conditions.</td>
</tr>
<tr>
<td>type of trajectory: initial MS BV-AA-MDS trajectory</td>
<td>An MS BV-AA-MDS trajectory generated on the 1st iteration of a multiple shooting method.</td>
</tr>
<tr>
<td>type of trajectory: IV-AA-MDS trajectory</td>
<td>A trajectory that satisfies conditions of an IVP. So, it satisfies initial conditions and also satisfies discretized Newtonian equations of motion.</td>
</tr>
<tr>
<td>type of trajectory: MS BV-AA-MDS solution trajectory</td>
<td>A BV-AA-MDS trajectory generated by a multiple shooting method.</td>
</tr>
<tr>
<td>type of trajectory: MS BV-AA-MDS trajectory</td>
<td>A trajectory generated by a multiple shooting method with $N$ subintervals that satisfies discretized Newtonian equations of motion on each of the $N$ subintervals.</td>
</tr>
<tr>
<td><em>well</em> $\rightarrow$ <em>well</em></td>
<td>A transition between two conformations belonging to different potential wells. If two potential wells are named (say $A$ and $B$), then the notation $A$ well $\rightarrow$ B well will refer to a transition from the potential well $A$ to the potential well $B$. (Informally, a potential well is a set of conformations that retain essential properties of the conformation of a local minimum of a potential energy surface. See section 2.2.2 for more details and a more formal characterization.)</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table 3.1  Summary data: local minima of $N$-acetyl-$N'$-methylalaninamide in vacuo .................. 157
Table 3.2  Initial trajectory generation, forward direction.............................................................. 157
Table 3.3  Initial trajectory generation, bidirectional, $C7_{eq}$ $\rightarrow$ $C7_{ax}$.................................................. 158
Table 3.4  Summary data: convergence for $min \rightarrow min$ BVP's .................................................. 158
Table 3.5  Example of convergence data for $min \rightarrow min$ BVP's .................................................. 159
Table 4.1  Indices of atom pairs for $N$-acetyl-$N'$-methylalaninamide.............................................. 204
Table 4.2  Classification of atom pairs for $N$-acetyl-$N'$-methylalaninamide................................. 205
Table 4.3  Summary of convergence for $C7_{ax}$ well $\rightarrow$ $C7_{eq}$ well BVP’s.......................................... 206
LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Ball-and-stick visualization of N-acetyl-N'-methylalaninamide</td>
</tr>
<tr>
<td>1.2</td>
<td>Ball-and-stick visualization: φ and ψ dihedral angles of N-acetyl-N'-methylalaninamide</td>
</tr>
<tr>
<td>1.3</td>
<td>φ-ψ contour plot: local minima of N-acetyl-N'-methylalaninamide</td>
</tr>
<tr>
<td>2.1</td>
<td>Depiction of components of an empirical potential energy function</td>
</tr>
<tr>
<td>2.2</td>
<td>Multiple shooting algorithm applied to Mueller potential</td>
</tr>
<tr>
<td>2.3</td>
<td>Condition numbers for N-acetyl-N'-methylalaninamide in vacuo: C7eq → C6</td>
</tr>
<tr>
<td>2.4</td>
<td>Condition numbers for N-acetyl-N'-methylalaninamide in vacuo: C7eq → C5p</td>
</tr>
<tr>
<td>2.5</td>
<td>Condition numbers for N-acetyl-N'-methylalaninamide in vacuo: C7ax → C7eq</td>
</tr>
<tr>
<td>2.6</td>
<td>Threshold time interval for well-conditioned single shooting Jacobian</td>
</tr>
<tr>
<td>2.7</td>
<td>Transition of a triangle: linear DMI</td>
</tr>
<tr>
<td>2.8</td>
<td>Transition of a triangle: nonlinear DMI</td>
</tr>
<tr>
<td>3.1</td>
<td>Ball-and-stick visualization of N-acetyl-N'-methylalaninamide</td>
</tr>
<tr>
<td>3.2</td>
<td>φ-ψ contour plot: local minima of N-acetyl-N'-methylalaninamide</td>
</tr>
<tr>
<td>3.3</td>
<td>Ball-and-stick visualization: local minima of N-acetyl-N'-methylalaninamide</td>
</tr>
<tr>
<td>3.4</td>
<td>BV-AA-MDS solution and three approximate BV-AA-MDS solutions</td>
</tr>
<tr>
<td>3.5</td>
<td>Strategic initial parameter vector selection for C7eq → C5p</td>
</tr>
<tr>
<td>3.6</td>
<td>Forward IVP's from C7eq local minimum; random initial velocity selection</td>
</tr>
<tr>
<td>3.7</td>
<td>Initial MS BV-AA-MDS trajectories, N = 3: mesh 1</td>
</tr>
<tr>
<td>3.8</td>
<td>Initial MS BV-AA-MDS trajectories, N = 3: mesh 2</td>
</tr>
<tr>
<td>3.9</td>
<td>Finding shooting points using a hidden time BV-AA-prx trajectory</td>
</tr>
<tr>
<td>3.10</td>
<td>Initial MS BV-AA-MDS trajectories, N = 3: mesh 2</td>
</tr>
<tr>
<td>3.11</td>
<td>Distribution of total energy in kcal mol⁻¹</td>
</tr>
<tr>
<td>3.12</td>
<td>Ensemble of trajectories</td>
</tr>
<tr>
<td>3.13</td>
<td>Initial and solution trajectories</td>
</tr>
<tr>
<td>4.1</td>
<td>φ-ψ contour plot: C7eq potential energy well of N-acetyl-N'-methylalaninamide</td>
</tr>
<tr>
<td>4.2</td>
<td>φ-ψ plots for low total energy simulations</td>
</tr>
<tr>
<td>4.3</td>
<td>Interatomic distance ranges by conformation type</td>
</tr>
<tr>
<td>4.4</td>
<td>φ-ψ plot of final trajectories</td>
</tr>
<tr>
<td>4.5</td>
<td>φ-ψ plot for initial trajectories: all modes</td>
</tr>
<tr>
<td>4.6</td>
<td>φ-ψ plot for solution trajectories: all modes</td>
</tr>
<tr>
<td>4.7</td>
<td>φ-ψ plot for initial trajectories: slowest 5 modes</td>
</tr>
<tr>
<td>4.8</td>
<td>φ-ψ plot for solution trajectories: slowest 5 modes</td>
</tr>
<tr>
<td>4.9</td>
<td>Distribution of total energy in kcal mol⁻¹</td>
</tr>
<tr>
<td>4.10</td>
<td>Vibrational modes of C7ax local minimum in cm⁻¹</td>
</tr>
<tr>
<td>5.1</td>
<td>Ball-and-stick visualization: three dihedral angles of N-acetyl-N'-methylalaninamide</td>
</tr>
<tr>
<td>5.2</td>
<td>φ/ψ/θ/DME/U plots: linear AA-DMI-ENM with uniform weights</td>
</tr>
<tr>
<td>5.3</td>
<td>φ/ψ/θ/DME/U plots: low energy solution trajectories</td>
</tr>
<tr>
<td>5.4</td>
<td>DDD plots: approximate MDS solution: C7eq → C6 transition</td>
</tr>
<tr>
<td>5.5</td>
<td>DDD plots: approximate MDS solution: C7eq → C5p transition</td>
</tr>
<tr>
<td>5.6</td>
<td>DDD plots: approximate MDS solution: C7eq → C7ax transition</td>
</tr>
<tr>
<td>5.7</td>
<td>DDD plots: approximate MDS solution II: C7eq → C7ax transition</td>
</tr>
<tr>
<td>5.8</td>
<td>DDD plots: MDS solution: C7eq → C7ax transition</td>
</tr>
<tr>
<td>5.9</td>
<td>DDD plots: linear DMI-ENM: C7eq → C7ax transition</td>
</tr>
<tr>
<td>5.10</td>
<td>φ/ψ/θ/DME/U plots: nonlinear AA-DMI-ENM with target distances from a BV-AA-MDS solution trajectory</td>
</tr>
<tr>
<td>5.11</td>
<td>φ/ψ/θ/DME/U plots: nonlinear AA-DMI-DG with target distances from a BV-AA-MDS solution trajectory</td>
</tr>
<tr>
<td>5.12</td>
<td>φ/ψ/θ/DME/U plots: C7eq → C7ax transition</td>
</tr>
<tr>
<td>5.13</td>
<td>DDD plots: overlay: C7eq → C7ax transition</td>
</tr>
<tr>
<td>5.14</td>
<td>DDD plots: zoom overlay: C7eq → C7ax transition</td>
</tr>
</tbody>
</table>
LABELING OF MATHEMATICAL EXPRESSIONS AND USAGE OF SYMBOLS

Note to the reader:

This dissertation includes many equations, inequalities, and other mathematical expressions. There are also definitions, theorems, algorithms, and hypotheses. These items have been labeled by chapter number followed by a period and then a number assigned by the order in which they appear within the chapter, all enclosed in parentheses. An effort has been made to avoid the use of symbols for multiple purposes. An exception is the use of the symbols in the component functions of the potential energy function $U(x)$ in equations (2.5) and (3.35). The component functions are shown for illustrative purposes. Symbols in these equations are used in a way that is common in literature. But, many of them also have other uses in this dissertation.
LIST OF REDUNDANCIES

Note to the reader:

This dissertation is divided into six chapters. The first chapter is an introductory chapter. The last chapter contains a summary and provides some possible directions for future research. The remainder has been organized into four chapters. It is anticipated that each of these four chapters will later be rearranged and tailored as an article for submission to a peer-reviewed journal. As a result, there are some redundancies between chapters similar to those that might be found in a collection of journal articles on different aspects of a common topic. There is some overlap when comparing the introductory sections of the different chapters. But, there is also some overlap between other sections as well. To help guide the reader of multiple chapters, a table of the primary redundancies is provided below:

<table>
<thead>
<tr>
<th>Heading number(s) of first extract</th>
<th>Heading number(s) of second extract</th>
<th>Heading number(s) of third extract</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.2</td>
<td>2.1-2.2</td>
<td>n/a</td>
<td>There is significant overlap in the material of these extracts. Neither extract is precisely a subset of the other, but of the common material, the second extract has more detail.</td>
</tr>
<tr>
<td>1.3</td>
<td>3.4.2</td>
<td>n/a</td>
<td></td>
</tr>
<tr>
<td>2.1.4-2.1.6</td>
<td>3.3.2-3.3.4</td>
<td>4.3.1-4.3.3</td>
<td></td>
</tr>
<tr>
<td>2.2.1</td>
<td>3.4.1</td>
<td>n/a</td>
<td></td>
</tr>
<tr>
<td>2.3.1-2.3.2</td>
<td>3.3.5-3.3.6</td>
<td>n/a</td>
<td>These extracts are very similar except that the first extract includes examples and the second does not.</td>
</tr>
<tr>
<td>2.3.4-2.3.6</td>
<td>3.3.8</td>
<td>n/a</td>
<td>The first extract is much more detailed.</td>
</tr>
<tr>
<td>2.3.8.2</td>
<td>3.2.1</td>
<td>n/a</td>
<td></td>
</tr>
<tr>
<td>2.3.10</td>
<td>5.2</td>
<td>n/a</td>
<td>The first extract contains examples; the second does not.</td>
</tr>
</tbody>
</table>
ABSTRACT

Determining the relationships between a protein’s structure and its function, identifying a protein’s function in biological systems, and understanding the mechanisms that facilitate a protein’s function are major scientific challenges in modern biological research. The mechanisms that facilitate protein function often involve conformational transitions of proteins or other biomolecules. Successfully meeting the aforementioned challenges will likely require an understanding of the dynamics of conformational transitions of proteins and of other biomolecules. Molecular dynamics is often studied using a computational approach that is called initial value all-atom molecular dynamics simulation (IV-AA-MDS) in this dissertation. In IV-AA-MDS, an empirically determined force field specifies the forces on each atom of a molecular system as a function of the coordinates of the system. The motions of the atoms of the system are governed by Newtonian equations of motion and are tracked dynamically over a period of time. From a mathematical perspective, in IV-AA-MDS, initial conditions for the positions and velocities of the atoms of the system are assigned for an initial point in time and the positions and velocities at a discrete set of points contained in an interval that includes the initial point in time are determined by using a numerical method for solving initial value problems (IVP’s) for systems of 2nd-order nonlinear ordinary differential equations (ODE’s).

The focus of the research of this dissertation is the mathematical modeling of and use of numerical methods for the study of the dynamics of conformational transitions of biomolecules like proteins and small peptides. While an IV-AA-MDS approach could be considered for this purpose, the focus of this dissertation is a related approach that is called boundary value all-atom molecular dynamics simulation (BV-AA-MDS) in this dissertation. This approach includes the application of a numerical method to seek numerical solutions to two-point boundary value problems (BVP’s) for systems of 2nd-order nonlinear ordinary differential equations (ODE’s). Numerical solutions to two-point BVP’s satisfy boundary conditions at two points in time (a beginning and ending point) and also satisfy Newtonian equations of motion at a discrete set of time points contained in the interval between the two points in time (within limits
of accuracy associated with the numerical method and the computer system being used). The numerical method of primary interest for this dissertation is multiple shooting.

For a brief description of multiple shooting, consider a molecular system with \( n \) atoms. And, assume that the time interval between the beginning and ending points is divided into \( N \) non-overlapping subintervals. If initial conditions for the positions and velocities of the atoms of the system are assigned for a point in time on each subinterval, then, on each subinterval, an IVP for a \( 3n \)-dimensional 2\(^{nd}\)-order differential equation can be formulated and a numerical solution can be obtained. Each of the \( N \) numerical solutions provides a \( 3n \)-dimensional trajectory for the positions of the \( n \) atoms and \( 3n \)-dimensional trajectory for the velocities of the \( n \) atoms. The \( N \) numerical solutions can be concatenated to form a trajectory for the entire time interval. A multiple shooting algorithm for a nonlinear BVP begins with a guess of what the initial conditions for the \( N \) IVP’s need to be so that the concatenated trajectory forms is a continuous trajectory that satisfies the boundary conditions. After the first iteration, typically, the \( 3n(N–1) \) position and \( 3n(N–1) \) velocity trajectories will have some jump discontinuities at the \( N–1 \) nodes dividing the subintervals and not all of the \( R \) boundary conditions will be satisfied. But, a system of \( 6n(N–1)+R \) nonlinear equations (NLE’s) can be derived. Iterative methods for solution of this system determine, on each iteration, adjustments to the initial positions and initial velocities for the IVP’s of the subintervals. On the next iteration, the adjusted initial data are used to solve the IVP’s. The solution of the NLE’s will correspond to a trajectory for which positions and velocities are continuous at each node and for which the BVP is satisfied. Single shooting is a special case of multiple shooting in which there is only one subinterval, i.e. \( N=1 \).

In this dissertation, the mathematical framework of AA-MDS, BV-AA-MDS and some numerical methods for BV-AA-MDS — single shooting, multiple shooting, finite differences methods, and stochastic difference equation methods — are described. Important computational limitations of AA-MDS, BV-AA-MDS, and MS for BV-AA-MDS are highlighted and reasons for considering these approaches and methods despite the computational limitations will be provided.
Also, in this dissertation, the application of multiple shooting to BVP’s for ODE’s corresponding to transitions between two molecular conformations specified by two sets of internal coordinates is proposed. Strategies and issues related to definition of boundary conditions, assignment of initial parameters, and convergence are investigated. Results from the study of transitions between local minima of the potential energy surface of an alanine dipeptide are presented. Implications of the methods and results of this work for application of multiple shooting to the study of conformational transitions in larger systems are discussed.

Defining boundary conditions corresponding to sets of internal coordinates of local minima leads to what is defined to be a full set of $6n$ boundary conditions, i.e. $R = 6n$. And, defining parameters of the multiple shooting method as the initial conditions on each subinterval leads to what is defined to be a full set of $6nN$ parameters. To apply multiple shooting with a full set of parameters to a BVP with a full set of boundary conditions, the number of atoms in the molecule must be limited to avoid excessive computational cost. In this dissertation, for the case of single shooting, an alternate boundary value simulation approach is presented that involves a reduced set of boundary conditions and a reduced set of parameters. BVP’s are constructed with boundary conditions defined as lower and upper bounds for selected interatomic distances that are intended to approximate potential energy wells. Modeling conformations transitions between potential energy wells has advantages in comparison with modeling conformational transitions between local minima of a potential energy surface. The former approach more closely reflects the reality of the physical problem being modeled and it allows for the possibility of a reasonably small reduced set of boundary conditions for a large system. A boundary condition can also be added to define bounds for the total energy of the system. We also propose an approach for use a reduced parameter set that is based on an application of principles of normal mode analysis. We provide results from the application of these approaches to the study of transitions between potential energy wells for an alanine dipeptide.

In this dissertation, all-atom distance matrix interpolation (AA-DMI) methods are described. These are methods for generating position trajectories that satisfy certain types of boundary conditions are
less computationally demanding than boundary value approaches to AA-MDS, but do provide atomically
detailed trajectories. These methods involve an optimization problem with an objective function derived by
interpolation of interatomic distances between their values in one conformation and their values in another
conformation. They can be expected to generate position trajectories that satisfy specified boundary
conditions, but do not necessarily satisfy Newtonian equations of motion. AA-DMI methods have practical
application in BV-AA-MDS as a means for generating initial trajectories for iterative methods like multiple
shooting. When BV-AA-MDS is applied, the range of values from which initial parameters for an iterative
numerical method must be selected in order to achieve eventual convergence is limited. So, selection of
appropriate initial parameters is important as likelihood of convergence can be impacted by the method
used for generating initial parameters. We consider conformational transitions in the alanine dipeptide,
\textit{N-acetyl-N'\textquotesingle-methylalaninamide}, and identify some of the difficulties with an all-atom version of a DMI
method that was previously introduced as an elastic network model (ENM). We introduce another
AA-DMI method based on ideas and methods commonly used in molecular distance geometry (DG) and
multidimensional scaling. We also propose the use of interpolation by spline functions as an alternative to
more the conventional and easily obtained interpolation by a linear polynomial. Refinement of AA-DMI
position trajectories by constrained energy minimization is also proposed. Results are presented from the
study of conformational transitions of an alanine dipeptide. Future directions of research are discussed.
1 OVERVIEW

1.1 Introduction

One major scientific challenge is to determine the relationships between a protein’s amino acid sequence and its structure and function. To fully comprehend these relationships, it likely requires that the dynamics of proteins and of other important types of biomolecules be understood. The focus of the research described in this dissertation is the development of methods for the study of the dynamics of conformational transitions of biomolecules like proteins and small peptides by molecular dynamics simulation (MDS). Knowledge of the dynamics of conformational transitions of proteins and other biomolecules is important for many areas of research in molecular biology — cell signaling, cell regulation, transport, pathology, drug discovery, functional site identification, protein-protein interactions, protein-small molecule interactions, protein design, and protein engineering. Simulation is a customary way to study molecular dynamics and MDS using an all-atom empirically determined force field is a well-established approach for the simulation of molecular processes. Conformational transitions can be studied by all-atom MDS within the mathematical framework of numerical solutions to nonlinear boundary value problems (BVP’s) for systems of 2nd-order ordinary differential equations (ODE’s). We will refer to this mathematical framework as boundary value all-atom MDS (BV-AA-MDS). To use this approach, starting and ending conformations of the molecule must be described mathematically.

Multiple shooting methods – a class of numerical methods for solving boundary value problems for ordinary differential equations – are applied here for isolated systems subject to an all-atom force field to find molecular dynamics trajectories that satisfy specified boundary conditions and satisfy Newton’s equations of motion. (An isolated system is a system with a fixed number of atoms, a fixed volume, and a fixed energy level.) The trajectories that satisfy the boundary conditions correspond to simulated conformational transitions of proteins. In general, important aspects of this work include assessing potential biological significance, comparing with experimental studies and computational studies, and
addressing computational issues. Computational issues include assessment of instability and its impact, assessment of feasibility of an approach for large systems, choice of global convergence schemes, development of efficient methods for choosing initial parameters, choice of boundary conditions, development of methods for parameter reduction, and assessment of algorithms for solving initial value problems, for computing Jacobian matrices, and for solving resulting nonlinear systems of equations.

The difficulties arising from computational challenges of BV-AA-MDS can be expected to increase significantly with the system size and duration of transition events. Even with use of a global convergence scheme for solving systems of nonlinear equations, in practice, the selection of the initial parameters can be critical for achieving convergence since the progress of global convergence schemes can be very slow until the Newton step can be used. The Newton step is usually only effective near a solution, so convergence within a reasonable number of iterations can only be expected if the initial parameters are sufficiently similar to parameters of a solution to the BVP.

In [Kim2002a], a description is provided of a distance matrix interpolation (DMI) method for generating trajectories that doesn’t necessarily satisfy Newton’s equations of motion, but does satisfy specified boundary conditions. All-atom DMI methods involve interpolation of interatomic distances between their values in one conformation and their values in another conformation. The method introduced in [Kim2002a] is based on a type of elastic network model and was coarse-grained in the sense that not all atoms were explicitly modeled. A typical coarse-grained approach to elastic network modeling would be to use one point for each amino acid, or residue. Applications of coarse-grained DMI can be found in [Kim2002b], [Kim2003], and [Kim2005]. With minor modifications, DMI methods can be applied at the atomic level of detail as well. In this dissertation, DMI approaches that incorporate methods commonly used in distance geometry, multidimensional scaling, and approximation of functions are introduced and applied at the all-atom level of resolution. They hold out the promise of efficient construction of initial trajectories as well as an alternative and possibly advantageous approach for the construction of trajectories that satisfy boundary conditions and approximately satisfy Newton’s equations of motion. The various
methods described in this dissertation have been applied to the study of conformational transitions of an alanine dipeptide and results will be presented.

1.2 Preliminaries

There is some common physical terminology associated with molecular conformations and transitions. The potential energy of a system is the energy inherent in the system due to the relative locations of the atoms within the physical system. It can be expected to vary as the relative location of the atoms of the system vary. The potential energy function for a molecule or system is a real valued function whose domain is the set of all conformations of the system. With the locations of the atoms in the system playing the role of variables, the value of the potential energy function can be represented as a curve, surface, or hypersurface. This surface is commonly known as the potential energy surface. In this dissertation, the conventional function name $U(x)$ will be used to refer to the potential energy function. The input to this function, $x$, is an $nd \times 1$ vector where $n$ is the number of particles in the system and $d \in \{1, 2, 3\}$ is the spatial dimension of the system. Simple but instructive examples can be realized by considering systems with only one particle and two spatial dimensions. For these types of examples, using rectangular coordinates with two dimensions, the conformation of the system can be represented graphically by a point in a plane and the potential energy surface can be represented graphically by a surface above the plane.

The scope of this dissertation with respect to potential energy functions, is limited to those that are twice-differentiable on their domain. The gradient of $U(x)$ is denoted by $\nabla U(x)$, and the Hessian of $U(x)$ is denoted by $\nabla^2 U(x)$.

There are some attributes of a potential energy function that will be of particular interest in this dissertation. A common distance measurement between two vectors is the Euclidean distance. The Euclidean distance is defined for two real-valued column vectors — $v$ and $w$ — with the same number of components as $\|v-w\|_2 = ((v-w)^T(v-w))^{1/2}$. A local minimizer of a potential energy function $U$ is a conformation, $x^*$, for which $U(x^*) \leq U(x)$ for all $x$ satisfying $\|x-x^*\|_2 < \epsilon$ for some $\epsilon > 0$. An isolated local minimizer of a potential function $U$ is a conformation, $x^*$ for which $U(x^*) < U(x)$ for all $x$ with
Also, we can say $U$ has a local minimum, $U(x^*)$, at $x^*$. Here, we will be primarily interested in isolated local minimizers and, by default, reference to a local minimizer should be understood to be reference to an isolated local minimizer. From optimization theory, we have that a necessary condition for $x^*$ to be a local minimizer is that $\nabla U(x^*) = 0$ ([Noc2002]).

Molecular modeling refers to the use of theoretical and computational methods and techniques to model the behavior of a molecule or system of molecules. The molecule or system of molecules can range from small to large. Molecular modeling based on quantum mechanics is beyond the scope of this dissertation. The most detailed modeling methods studied here will be based on Newtonian, or classical, mechanics. Some justification for this scope will be provided in section 2.3.8.1. All-atom molecular dynamics simulation (AA-MDS) generally refers to a particular type of molecular modeling in which the motion of the atoms or particles of the molecules of the system are tracked dynamically over a period of time and the motion is governed deterministically by Newtonian equations of motion. More specifically,

\begin{equation}
M \mathbf{a}(t) = \mathbf{f}(\mathbf{x}(t)),
\end{equation}

where $t$ is a scalar representing time, $t_0$ represents the beginning time, $t_f$ represents the ending time, $\mathbf{x}(t)$, $\mathbf{v}(t)$, and $\mathbf{a}(t)$ are $3n \times 1$ vectors representing the position, velocity, and acceleration, respectively, of the $n$ particles of the system at time $t$ in three dimensions of a rectangular coordinate system, $M$ is a $3n \times 3n$ diagonal matrix with the mass of each particle repeated in the three diagonal entries corresponding to the 3 dimensions of physical space, and $\mathbf{f}(\mathbf{x}(t))$ is an $3n \times 1$ vector representing the force acting on each particle of the system at time $t$ in each dimension. Note that $\mathbf{v}(t) = \mathbf{x}'(t)$ and $\mathbf{a}(t) = \mathbf{x}''(t)$, so (1.1) is a 2nd-order ordinary differential equation (ODE). A more general representation of $\mathbf{f}$ would give $\mathbf{f}$ as a function of both $t$ and $\mathbf{x}(t)$, i.e. $\mathbf{f}(t, \mathbf{x}(t))$. Since $\mathbf{f}$ is not an explicit function of $t$, we say that (1.1) is an autonomous ODE. The function, $\mathbf{f}(\mathbf{x}(t))$ may be obtained by theoretical or empirical means. In this dissertation, we assume that there are no stochastic or random terms in $\mathbf{f}(\mathbf{x}(t))$. In order to begin a simulation, additional specifications are required.

For IV-AA-MDS, additional specifications are the initial values of the form
where $\dot{x}$ and $\dot{v}$ are $3n \times 1$ vectors and $t_0 \leq t \leq t_f$. Equations (1.1) and (1.2) define an initial value problem (IVP).

For $f$ linear, the domain for existence and uniqueness of solutions can be specified by inspection of $f$. For $f$ nonlinear, the entire domain cannot be specified. For two-point boundary value (BV-) AA-MDS, additional specifications are given by

$$
(1.3) \quad r(\dot{x}(t_0), \dot{v}(t_0), \dot{x}(t_f), \dot{v}(t_f)) = 0
$$

where $r$ is an $R \times 1$ vector for some integer $R$. Equations (1.1) and (1.3) define a two-point boundary value problem (BVP). For BVP’s of this form, in general, there may or may not be a solution, and if there is a solution, it may not be unique. The adjective ‘two-point’ indicates that $r$ is a function describing the characteristics of the system at two points in time, $t_0$ and $t_f$. BVP’s with boundary conditions at more than two points are called multipoint BVP’s. We will be focusing on two-point BVP’s in this dissertation, so reference to a BVP will, by default, be a reference to a two-point BVP. The function $f$ is commonly called a force field. If $f(x)$ is the gradient of a real-valued function of $x$, then $f(x)$ is a conservative force field, and there exists a potential energy function, $U(x)$, that satisfies $f(x) = -\nabla U(x)$.

It is possible that a solution to a specific IVP or BVP may be obtained by analytical means. But, the IVP’s and BVP’s that arise in AA-MDS will, in general, require use of a numerical method implemented on a computer to obtain an approximate solution. The solution is only approximate due to error in the numerical method and error in arithmetic calculations performed in finite machine arithmetic. The term position trajectory, refers to an ordered set of coordinates assumed by a dynamical system on a discrete mesh of time points. The term velocity trajectory refers to an ordered set of velocities corresponding to a position trajectory. A trajectory consists of a position trajectory and a velocity trajectory. It can be directly defined as an ordered set of intermediate states assumed by a dynamical system on a discrete mesh of time points. If there are $\mathcal{D}$ increments in time from $t = t_0$ to $t = t_f$, then the mesh will have $\mathcal{D} + 1$ intermediate states (including the beginning and ending states). The $i^{th}$ intermediate state of the trajectory is known as the $i^{th}$ snapshot of the trajectory. For a position or velocity trajectory
with $D+1$ snapshots including the endpoints, the trajectory will be represented by a $3n \times (D+1)$ matrix. So, columns of this matrix correspond to snapshots of the trajectory. Rows correspond to the evolution of the position or velocity of one particle in one coordinate direction. Using an appropriate force field, a sufficiently accurate numerical method and an adequately small time step, AA-MDS can be applied to generate physically meaningful trajectories.

To apply numerical methods for solution of BVP’s to BV-AA-MDS, one must define boundary conditions and determine an appropriate length of time for the simulation. With respect to boundary conditions for the study of conformational transitions, note that important molecular conformations are often associated with a particular local minimum of the potential energy function, or an equivalence class of local minima of the potential energy function. Boundary conditions can be defined to correspond to the relevant local minima or equivalence class of local minima. Different variations on this approach will be described in Chapter 3. In addition, Chapter 3 will include an approach for assigning an appropriate length of time for simulation.

To model conformational transitions by trajectories between local minima is convenient mathematically. It should be noted, however, that, from a biological perspective, a conformation of a molecule may be classified as being in the conformation of a particular local minimum as long as it is approximately in that conformation. Both in reality and in the realistic theoretical application of Newtonian physics to an isolated system, the system will contain some energy and due to that energy, a molecule that is, from a practical perspective, in a certain conformation associated with a local minimum would not remain in precisely in that conformation, but would be subject to constant vibrational or oscillatory motion ‘around’ that local minimum. In fact, the precise conformation of the local minimum itself would be one of many possible conformations and may actually rarely be realized. This local motion may not affect the overall conformation from a qualitative or practical perspective. A more realistic model construction of boundary conditions might take into account this observation that molecules are in constant motion. Boundary conditions might be defined so that they would be satisfied as long as starting and ending conformations are appropriately ‘near’ the relevant local minima. More precision with respect to
the usage of the terms ‘around’ and ‘near’ above will be provided in subsection 2.2.2 and in Chapter 4. For now, we note that a conformation ‘around’ or ‘near’ a local minimum is a conformation that retains essential properties of the conformation of the local minimum. And, the set of all conformations satisfying this properties for a local minimum with name ‘A’ will be called the ‘A’ potential well, or ‘A’ well.

Multiple shooting is a numerical method for solving BVP's for ODE's. For introductory purposes, a brief description is given here. (The details of the application of this method to the BVP (1.1), (1.3) will be presented in Chapter 2.) The time interval of the BVP is divided into $N$ non-overlapping subintervals. It is convenient to first convert the $3n$-dimensional 2nd-order differential equation (1.1) to a $6n$-dimensional 1st-order differential equation. An initial value problem (IVP) for this $6n$-dimensional 1st-order differential equation is numerically solved on each of the $N$ non-overlapping subintervals of the time interval of the simulation. On each subinterval, a numerical solution to the IVP provides a $3n$-dimensional trajectory for the positions of the $n$ atoms and $3n$-dimensional trajectory for the velocities of the $n$ atoms. The $N$ numerical solutions can be concatenated to form a trajectory for the entire time interval. The algorithm begins with a guess of what the initial conditions for the IVP’s need to be so that these trajectories can be concatenated to form a continuous trajectory that satisfies the boundary conditions. After the first iteration, typically, the $6nN$ position and velocity trajectories will have some jump discontinuities at the $N-1$ nodes dividing the subintervals and not all of the $R$ boundary conditions will be satisfied. But, a system of $6n(N-1)+R$ nonlinear equations (NLE’s) can be derived. Iterative methods for solution of this system determine, on each iteration, adjustments to the initial positions and initial velocities for the IVP’s of the subintervals. On the next iteration, the adjusted initial data are used to solve the IVP’s. The solution of the NLE’s will correspond to a trajectory for which positions and velocities are continuous at each node and for which the BVP is satisfied. This approach has a natural extension to a parallel or distributed environment since a multiple shooting algorithm can be defined so that the IVP trajectories on the different subintervals can be obtained on separate processors. Defining boundary conditions corresponding to coordinates of local minima leads to $6n$ boundary conditions, i.e. $R = 6n$. And, defining parameters of the multiple shooting method as the initial conditions on each subinterval leads to $6nN$ parameters. For application to
large systems, we anticipate that it will be necessary to construct a relevant boundary value problem with $R$ significantly less than $6n$ (i.e. $R \ll 6n$) and that the number of parameters, $S$, required to determine the initial positions and velocities on the $N$ subintervals must be significantly less than $6nN$ (i.e. $S \ll 6nN$).

Methods for attempting to solve the nonlinear system of equations involve, at each iteration, solving a linear system of complexity $O(Nn^3)$. This notation indicates that number of numerical operations required to solve the linear system is bounded by a function $\tilde{z}(N,n) = CNn^3$, for some real positive constant $C$ as $N, n \to \infty$. An approach that has the potential for solving BVP’s corresponding to conformational changes in proteins and reduces this complexity to an order potentially lower than $O(Nn^3)$ for the case $N=1$ is proposed in Chapter 4. It is possible that a similar approach could be developed for $N>1$.

Taking a chronological and conceptual step backward, it is noted that the author began considering the possibility of boundary value all-atom molecular dynamics simulations (BV-AA-MDS) and specifically the application of the multiple shooting method as a thesis topic at the beginning of the second year of graduate study. Time was spent acquiring the necessary background knowledge on numerical methods for solving BVP’s, molecular dynamics simulation, globally convergent approaches to solution of nonlinear equations, BV-AA-MDS, folding and conformational transitions of proteins, and other related topics. Experiments were performed in MATLAB with one-particle systems in one or two dimensions using Lennard-Jones potentials, double well potentials, the Mueller potential ([Ole1996], [Bai2006]) and other small model potentials using the multiple shooting method and other numerical methods such as the collocation method, and the stochastic difference equation method ([Elb1999]) for solving BVP’s. Also, in MATLAB, for a six particle Lennard-Jones cluster in three dimensions, simulation was used to find trajectories that transition between the global minimum, an octahedral structure, and a low-energy local minimum, a tripyramidal structure ([Hoa1971]). Furthermore, the BV-AA-MDS software, MOIL, ([Elb1994]) was used to simulate conformational transitions for small peptides and also a model protein using stochastic difference equation.

A natural question after this preliminary work was ‘Is further research on the application of the multiple shooting method for BV-AA-MDS justified?’ After introduction of appropriate background
material in the early part of Chapter 2, some justification of BV-AA-MDS as a means to study conformational transitions will be provided in section 2.4.2. And, in section 2.4.3, some justification of the use of the multiple shooting method as a numerical method for BV-AA-MDS will be provided. Some general comments that are intended to help justify further study are also provided in the remainder of this paragraph. In the realm of BV-AA-MDS, there is a lesser amount of reported work associated with the multiple shooting method than least-action and finite difference methods such as the stochastic difference equation method and related methods. (These methods will be described in Chapter 2.) If it is indeed reasonable to use the multiple shooting method, the relative scarcity of reported may indicated a greater potential for scientific gain with the multiple shooting method. Also, its natural extension to efficient parallel or distributed computing would enable applications to larger systems. Furthermore, with respect to the stochastic difference equation method, its customary use has been for the purpose of generating approximate trajectories over long time intervals, computationally unreachable by ‘exact’ (i.e. exact within the limits of numerical error) methods. While there may be effective ways to use the multiple shooting method for the same purpose, the primary emphasis of my work to date is its use as an alternative to initial value based approaches for generation of ‘exact’ trajectories. The stochastic difference equation method and other finite-difference methods could also be used for this purpose. There are advantages and disadvantages to each method. With this in mind, the study of multiple shooting approaches for BV-AA-MDS can be seen as complementary to current numerical approaches to BV-AA-MDS. Finally, the method is intriguing in that it can be considered to be, from an algebraic standpoint, a simple approach, but yet, from a geometric standpoint, it is somewhat surprising that one can find solution trajectories in high dimensions by this iterative procedure.

It is also noted that the MATLAB computing environment was convenient and appropriate for initial experiments. After the initial experiments, a goal for further work for this dissertation was to apply multiple shooting methods to the study of the motions representing conformational transitions of small peptides subject to a biomolecular all-atom force field. Since publicly available MATLAB implementations of biomolecular all-atom force fields appear to be lacking, the author developed a
MATLAB implementation of the AMBER all-atom potential energy function. The gradient and Hessian were also determined along with a variety of other useful and important functions for BV-AA-MDS. The force field can be immediately determined from the gradient. It is intended that MATLAB toolboxes will be created for molecular dynamics simulations and multiple shooting and it is conjectured that they should be useful for others in both teaching and research.

1.3 Overview of doctoral research

The methods described in this dissertation have been applied to \textit{N-acetyl-N'}-methylalaninamide, a twenty-two atom molecule commonly termed as an alanine dipeptide, in \textit{vacuo}. This molecule contains an amino acid with an alanine side chain—the C\text{\textgreek{b}} atom has three hydrogen atoms bonded to it, so the side chain is just a methyl group, CH\textsubscript{3}—capped with an acetyl group at the N-terminus and with an amide and a methyl group on the C-terminus (see Figure 1.1). The force field is the previously mentioned MATLAB implementation of the AMBER all-atom force field which is also implemented in the software MOIL with some variations. Two measurements — the C-N-C\text{\textalpha}-C dihedral angle (\textgreek{phi}) and the N-C\text{\textalpha}-C-N dihedral angle (\textgreek{psi}) — are termed ‘soft’ degrees of freedom for this molecule and are of primary importance in determining the overall shape of the alanine dipeptide. These dihedral angles are identified in Figure 1.2. For the temperature and environment of interest, most of the other internal degrees of freedom deviate only slightly from mean values as a function of time or are not influential in determining the overall shape. Because of the relative flexibility and the importance of the \textgreek{phi} and \textgreek{psi} dihedral angles, it is common to use a projection onto a two-dimensional subspace determined by the values of \textgreek{phi} and \textgreek{psi} to visualize the potential energy surface and also as a way to visualize conformational changes. A two dimensional adiabatic energy map for these two angles, constructed in MATLAB by constrained energy minimization, is shown in Figure 1.2. This figure includes the identification of the \textgreek{phi} and \textgreek{psi} values for six common local minima. This energy map suggests that at least four of the local minima of the potential energy surface — C\textsubscript{7\text{eq}}, C\textsubscript{6}, C\textsubscript{5\textgreek{beta}}, and C\textsubscript{7\text{ax}} — represent minima on this adiabatic energy map.
Figure 1.1  Ball-and-stick visualization of \textit{N-acetyl-N'}-methylalaninamide

\textit{Ball-and-stick visualization of N-acetyl-N'}-methylalaninamide Atoms are shown as balls; covalent bonds are shown as sticks. The colors of the atoms represent the type of atom. (hydrogen—white; carbon—greenish-gray; nitrogen—blue; oxygen—red).
**Figure 1.2** Ball-and-stick visualization: $\phi$ and $\psi$ dihedral angles of $N$-acetyl-$N'$-methylalaninamide

Ball-and-stick visualization: $\phi$, $\psi$, and $\theta$ dihedral angles of $N$-acetyl-$N'$-methylalaninamide. Figure 1 of [Bol2000]. The C-N-C$^\alpha$-C backbone dihedral angle ($\phi$) and the N-C$^\alpha$-C-N backbone dihedral angle ($\psi$) are labeled.

**Figure 1.3** $\phi$-$\psi$ contour plot: local minima of $N$-acetyl-$N'$-methylalaninamide

$\phi$-$\psi$ contour plot: local minima of $N$-acetyl-$N'$-methylalaninamide Marked in the $\phi$-$\psi$ contour plot are six primary local minima of the alanine dipeptide potential energy surface.
The research of this dissertation includes (1) development and implementation of multiple-shooting algorithms for BV-AA-MDS, (2) assessment of some computational limitations of BV-AA-MDS, (3) understanding the multiple shooting method in the context of other numerical methods used in current related research, (4) discovery and analysis of trajectories corresponding to conformational transitions between local minima, (5) discovery and analysis of trajectories that correspond to conformational transitions between the wells surrounding local minima of the potential energy surface, and (6) development and analysis of distance matrix interpolation methods and the application of these methods to generation of initial trajectories for multiple shooting methods for BV-AA-MDS.

1. **Development and implementation of multiple shooting for BV-AA-MDS**

   General descriptions of multiple shooting can be found in [Sto2002] and [Asc1995]. In Chapter 2 of this dissertation, a brief description of multiple shooting for application to BV-AA-MDS, is given.

2. **Assessment of some computational limitations of BV-AA-MDS**

   Limitations of applicability of different BVP methods are discussed in Chapter 2. Included are discussions of stability, and practical limits for the length of time interval of simulation due to limits on the step size for numerical solution of IVP’s. Additionally, limitations related to the number of variables and the number of boundary conditions are discussed in Chapter 2.

3. **Multiple shooting methods and other methods used in current related research**

   In Chapter 2, some perspective is provided by considering differences and similarities between the multiple shooting (MS) methods and finite difference methods for BV-AA-MDS. Similarities between a particular finite difference method for BV-AA-MDS called the stochastic difference equation (SDE) method ([Elb1999]) and a standard finite difference method are identified. Differences and similarities between the MS and SDE methods are highlighted by presentation of a simple example. Also, an approach for comparing the computational efficiency of IV-AA-MDS and multiple shooting methods for BV-AA-MDS is presented in Chapter 2.

4. **Conformational transitions between local minima for N-acetyl-N'-methylalaninamide**
For a simulation of the alanine dipeptide in vacuo with energy conservation, a transition from one local minimum to another is an extremely rare event. In general, the study of rare transitions can be of intrinsic interest. Also, the study of transitions between local minima can serve as a model for the study of transitions from one state to another state in larger systems. The notation \( \text{min} \rightarrow \text{min} \) will be used to refer to a general transition between local minima of a potential energy surface in this dissertation. If two local minima are named, say \( A \) and \( B \), then the notation \( A \rightarrow B \) will refer to a transition from local minimum \( A \) to local minimum \( B \). Application of multiple shooting to a \( \text{min} \rightarrow \text{min} \) BVP – that is a BVP in which boundary conditions are specified to correspond to local minima – requires a full set of \( 6n \) boundary conditions (\( 3n \) for both the beginning and ending conformations of the simulation), and generally requires a full set of \( 6nN \) parameters (\( 3n \) for both the initial positions and the initial velocities of the \( n \) atoms in the system on each of the \( N \) subintervals). With a full set of parameters, the multiple shooting algorithm for BV-AA-MDS exhibits a rate of convergence which seems to be superlinear, but convergence requires appropriate selection of initial parameters. In Chapter 3 of this dissertation, we describe in detail the multiple shooting algorithm and how this algorithm fits into a general computational strategy for generating a set of initial parameters. Also, in Chapter 3, we focus on three different transitions — \( C7_{eq} \rightarrow C6 \), \( C7_{eq} \rightarrow C5_{\beta} \), and \( C7_{eq} \rightarrow C7_{ax} \) — between local minima. We illustrate the convergence results for various numbers of subintervals, i.e. different values of \( N \).

5. **Conformational transitions between potential energy wells for N-acetyl-N'-methylalaninamide**

For larger and more complex motions in molecules like proteins, it is more practical and, perhaps, of greater interest to study the problem of conformational transitions corresponding to movement between wells that surround different local minima of a potential energy surface. Transitions between these wells, which will be called potential energy wells and will be more carefully characterized in Chapter 2, can be studied within the BVP approach by use of appropriately defined boundary conditions. The notation \( \text{well} \rightarrow \text{well} \) will be used to refer to a general transition between two conformations belonging to different wells of the potential energy surface. If two
wells are named, say $A$ and $B$, then the notation $A$ well $\rightarrow B$ well will refer to a transition from the well $A$ to the well $B$. In Chapter 4 of this dissertation, a method is developed for defining boundary conditions corresponding to potential energy wells that is based on bounds for selected interatomic distances and also an upper bound for total energy. This approach provides for the possibility of a relatively easily defined and reasonably small set of boundary conditions for large systems. If a full set of parameters is used, the number of MS parameters can become excessive. A full parameter set for the multiple shooting approach scales linearly with the product of the number of particles in the system and the number of multiple shooting subintervals. For the special case of a single multiple shooting subinterval (i.e. single shooting), we propose an normal-mode-based approach for significantly reducing the number of parameters. A reduced parameter set would be important for the application of shooting methods for large complex systems. It should be possible to develop a similar approach for application with multiple subintervals, and such an approach will be considered in work beyond this dissertation. In Chapter 4 of this dissertation, we provide results from the application of the approaches to reduce the number of boundary conditions and the number of parameters for single shooting to the study of well $\rightarrow$ well transitions for the alanine dipeptide.

6. **Distance matrix interpolation between local minima for N-acetyl-N'-methylalaninamide**

As previously mentioned, practical convergence of global convergence schemes for nonlinear equations is strongly dependent on the initial trajectory. There are methods to generate trajectories that are thought to be less accurate, and possibly less detailed, than AA-MDS. Included in this category are distance matrix interpolation (DMI) methods. These methods are useful for efficient construction of initial trajectories. A brief introduction to DMI will be provided in section 2.3.10. DMI methods involve interpolation of interatomic distances between values for beginning conformations and values for ending conformations. In Chapter 5 of this dissertation, we will describe previously developed DMI methods from a slightly different, but hopefully useful perspective and introduce ‘new’ methods that are based on the recognition that subproblems in the
generation of a AA-DMI position trajectory are equivalent to optimization problems that arise in the fields of molecular distance geometry and multidimensional scaling. We consider the example of conformational transitions in the alanine dipeptide. DMI position trajectories can sometimes include sequences of conformations with extremely high energy. A description is provided of how potential energy level of intermediate conformations can be controlled by performing constrained energy minimization on intermediate conformations for the alanine dipeptide.

1.4 References


2 BOUNDARY VALUE BIOMOLECULAR DYNAMICS SIMULATION AND NUMERICAL METHODS

This chapter will include a description of all-atom molecular dynamics simulation (AA-MDS) and a short description of some numerical methods—single shooting methods, multiple shooting methods, finite differences methods, and stochastic difference equation methods—for boundary value AA-MDS. While multiple shooting (MS) methods will be the numerical method that is emphasized in this dissertation, these other methods are introduced to facilitate some comparisons and to motivate the use of multiple shooting methods. The chapter will also include a description of two globally convergent optimization methods. Both methods are modifications of Newton’s method for solving nonlinear systems of equations and the use of these particular approaches will also be motivated. Important computational limitations of AA-MDS will be highlighted. Reasons for undertaking AA-MDS, BV-AA-MDS, and MS for BV-AA-MDS despite the computational limitations will be provided.

2.1 Preliminaries I

In this section, some important mathematical and physical concepts and terminology that are associated with the computational study of molecular conformations and transitions and all-atom molecular dynamics simulation are briefly introduced.

2.1.1 Potential energy functions

The potential energy of a system is the energy inherent in the system due to the relative locations of the atoms within the physical system. It can be expected to vary as the relative location of the atoms of the system vary. The potential energy function for a molecule or system is a real valued function whose domain is the set of all conformations of the system. With the locations of the atoms in the system playing the role of variables, the value of the potential energy function can be represented as a curve, surface, or hypersurface. This surface is commonly known as the potential energy surface. In this dissertation, the
conventional function name $U(x)$ will be used to refer to the potential energy function and the units, unless otherwise indicated, will be assumed to be $\text{keal mol}^{-1}$. The input to this function, $x$, is an $n \times d$-dimensional vector where $n$ is the number of particles in the system and $d \in \{1, 2, 3\}$ is the spatial dimension of the system. Simple but instructive examples can be realized by considering systems with only one particle and two spatial dimensions. For these examples, using rectangular coordinates, the conformation of the system can be represented graphically by a point in a plane and the potential energy surface can be represented graphically by a surface above the plane. Here, we will be primarily interested in potential energy functions that are twice-differentiable on their domain. The gradient of $U(x)$ is denoted by $\nabla U(x)$, and the Hessian of $U(x)$ is denoted by $\nabla^2 U(x)$.

2.1.2 Metrics

A function which satisfies some elementary properties desirable for measurement of distance between two vectors is known as a metric. A common metric is the Euclidean metric. The measure of distance for two real-valued column vectors — $v$ and $w$ — with the same number of components using this metric is $||v-w||_2 = ((v-w)^T(v-w))^{1/2}$. There are other metrics used to measure distance between vectors. An arbitrary metric will be indicated by $||\cdot||$.

2.1.3 Local minima

A local minimizer of a potential energy function $U$ is a conformation, $x^*$, for which $U(x^*) \leq U(x)$ for all $x$ satisfying $||x-x^*|| \leq \epsilon$ for some $\epsilon > 0$. An isolated local minimizer of a potential function $U$ is a conformation, $x^*$, for which $U(x^*) < U(x)$ for all $x$ with $0 < ||x-x^*|| \leq \epsilon$. Also, we can say $U$ has a local minimum, $U(x^*)$, at $x^*$. Here, we will be primarily interested in isolated local minimizers and by default reference to a local minimizer should be understood to be reference to an isolated local minimizer. From optimization theory, we have that a necessary condition for $x^*$ to be a local minimizer is that $\nabla U(x^*) = 0 ([\text{Noc2002}])$. 
2.1.4 All-atom molecular dynamics simulation (AA-MDS)

Molecular modeling is a general term referring to the use of theoretical and computational methods and techniques to model the behavior of a molecule or system of molecules. The molecule or system of molecules can range from small to large. All-atom molecular dynamics simulation (AA-MDS) generally refers to a particular type of molecular modeling in which the motion of the atoms or particles of the molecules of the system are tracked dynamically over a period of time and the motion is governed deterministically by the Newtonian equations of motion. In this dissertation, the scope will be limited to AA-MDS for an isolated system. An isolated system is a system with a fixed number of particles, a fixed volume, and a fixed amount of energy. In this case, the forces on each atom are determined by the arrangement of the atoms and are not explicitly dependent on time. In other words, the Newtonian system of differential equations are autonomous. More specifically,

\[ M \frac{d}{dt} a(t) = f(x(t)), \quad t_0 < t < t_f \]

where \( t \) is a scalar representing time, \( t_0 \) represents the beginning time, \( t_f \) represents the ending time, \( x(t), v(t), \) and \( a(t) \) are \( 3n \times 1 \) vectors representing the position, velocity, and acceleration, respectively, of the \( n \) particles of the system at time \( t \) in three dimensions of a rectangular coordinate system, \( M \) is a \( 3n \times 3n \) diagonal matrix with the mass of each particle repeated in the three diagonal entries corresponding to the 3 dimensions of physical space, and \( f(x(t)) \) is an \( 3n \times 1 \) vector representing the force acting on each particle of the system at time \( t \) in each dimension. Note that \( v(t) = x'(t) \) and \( a(t) = x''(t) \), so (2.1) is a 2nd-order ordinary differential equation (ODE). A more general representation of \( f \) would give \( f \) as a function of both \( t \) and \( x(t) \), i.e. \( f(t,x(t)) \). Since \( f \) is not an explicit function of \( t \), we say that (2.1) is an autonomous ODE. The function, \( f(x(t)) \) may be obtained by theoretical or empirical means. In this dissertation, we assume that there are no stochastic or random terms in \( f(x(t)) \). In order to begin a simulation, additional specifications are required.

2.1.5 Initial value AA-MDS (IV- AA-MDS)

For IV-AA-MDS, additional specifications are the initial values of the form
\[ (2.2) \quad x(t) = x' \quad \text{and} \quad v(t) = v' \]

where \( x' \) and \( v' \) are \( 3n \times 1 \) vectors and \( t_0 \leq \tau \leq t_f \). Equations (2.1) and (2.2) define an initial value problem (IVP). For \( f \) linear, the domain for existence and uniqueness of solutions can be specified by inspection of \( f \). For \( f \) nonlinear, the entire domain cannot be specified.

### 2.1.6 Boundary value AA-MDS (BV- AA-MDS)

For two-point boundary value (BV-) AA-MDS, additional specifications are given by

\[ (2.3) \quad r(x(t_0), v(t_0), x(t_f), v(t_f)) = 0 \]

where \( r \) is an \( R \times 1 \) vector for some integer \( R \). In this chapter and in Chapters 3 and 5, we will assume \( R = 6n \) and that the \( 6n \) boundary conditions can be separated into \( 3n \) conditions on \( x(t_0) \) and \( 3n \) conditions on \( x(t_f) \). So, the number of scalar boundary conditions is equal to the number of scalar differential equations in (2.1). In Chapter 4, we will consider BVP’s with \( R < 6n \). Equations (2.1) and (2.3) define a two-point boundary value problem (BVP). For a BVP of this form, in general, there may or may not be a solution, and if there is one solution, it might not be the only one. The adjective ‘two-point’ indicates that \( r \) is a function describing the characteristics of the system at two points in time, \( t_0 \) and \( t_f \). BVP’s with boundary conditions at more than two points are called multipoint BVP’s. We will be focusing on two-point BVP’s in this dissertation, so reference to a BVP will, by default, be a reference to a two-point BVP. The function \( f \) is commonly called a force field. If \( f(x) \) is the gradient of a real-valued function of \( x \), then \( f(x) \) is a conservative force field, and there exists a potential energy function, \( U(x) \), that satisfies \( f(x) = -\nabla U(x) \).

### 2.1.7 Generalized coordinates; relating potential energy functions and AA-MDS

It is possible that a solution to a specific IVP or BVP may be obtained by analytical means. But, the IVP’s and BVP’s that arise in AA-MDS will, in general, require use of a numerical method implemented on a computer to obtain an approximate solution. The solution is only approximate due to
error in the numerical method and error in arithmetic calculations performed in finite machine arithmetic.

The term *position trajectory*, refers to an ordered set of coordinates assumed by a dynamical system as a result of time. The term *velocity trajectory* refers to an ordered set of velocities corresponding to position trajectory of a dynamical system as a result of time evolution. A *trajectory* consists of a position trajectory and a velocity trajectory. It can be directly defined as an ordered set of intermediate states assumed by a dynamical system as a result of time. The $i^{th}$ intermediate state of the trajectory is known as the $i^{th}$ *snapshot* of the trajectory. For a position or velocity trajectory with $D+1$ snapshots including the endpoints, the trajectory will be represented by a $3n \times (D+1)$ matrix. So, columns of this matrix correspond to snapshots of the trajectory. Rows correspond to the evolution of the position or velocity of one particle in one coordinate direction. Using an appropriate force field, a sufficiently accurate numerical method and an adequately small time step, AA-MDS can be applied to generate physically meaningful trajectories.

To apply numerical methods for solution of BVP’s to BV-AA-MDS, one must define boundary conditions and determine an appropriate length of time for the simulation. With respect to boundary conditions for the study of conformational transitions, note that important molecular conformations are often associated with a particular local minimum of the potential energy function, or an equivalence class of local minima of the potential energy function. Boundary conditions can be defined to correspond to the relevant local minima or equivalence class of local minima. Different variations on this approach will be described in subsection 2.2.1. In addition, Chapter 3 will include an approach for assigning an appropriate length of time for simulation.

The potential energy of the system is related to Newton’s equations of motion by

$$f(x(t)) = -\nabla U(x(t)).$$

The form of the potential energy function that will be used primarily in this dissertation is given by:
where $K_b$ represents a bond-stretching force constant; $b$, a bond length; $b_0$, an ideal bond length; $K_\theta$, a bond-angle-bending force constant; $\theta$, a bond angle; $\theta_0$, an ideal bond length; $K_\phi$, a dihedral-angle-bending force constant; $\phi$, a dihedral angle; $n$, a dihedral angle multiplicity term; $\delta$, a phase factor; $A_{i,j}$, a van-der-Waals repulsion parameter; $B_{i,j}$, a van-der-Waals attraction parameter; $r_{i,j}$, an interatomic distance; $q_i$, an atomic electrostatic charge; $\varepsilon$, the dielectric constant; $v_{1-4}$, a van-der-Waals adjustment factor for 1-4 atom pairs; and $l_{1-4}$, an electrostatic adjustment factor for 1-4 atom pairs. In Figure 2.1, these different components are depicted using a simplified representations of atoms and two dimensional plots. More details are the components are included in the text below this figure. In (2.5), the argument to $U$ is a vector containing the rectangular coordinates of the atoms of the molecule. Given this argument, the bond lengths, bond angles, dihedral angles, and interatomic distances can readily be determined. But, none of the components of $U$ depends on the precise location the molecule; they all depend on relative locations of the atoms. It can be expected, then, that the local minima of $U$ depend only on relative locations of the atoms. Put another way, any rotation and/or translation of a local minimum is also a local minimum. The relative locations of atoms can be determined by internal coordinates, which are usually comprised of bond lengths, bond angles and dihedral angles. Systematic approaches for describing atomic structures using internal coordinates and conversion between internal and rectangular coordinates can be developed based on calculations given in [Tho1967], [Nik1977], [WuZ2003], [Elb2003], [Phi1995], and Chapter 1 of [Lea2001]. The internal coordinates are one type of what is commonly known as generalized coordinates. Generalized coordinates refer to essentially any coordinate system other than a rectangular coordinate system.
2.2 Ideas, Methods, and Analysis I

2.2.1 Specification of boundary conditions for AA-MDS

Boundary conditions can be fairly easily defined in terms of absolute position in three dimensional space. A conformation of a molecule is determined by relative locations of atoms, or equivalently, by internal coordinates. A solution to a BVP with boundary conditions defined in terms of absolute position requires, then, not only that the molecule changes from one conformation to another over the specified time interval, but also that the molecule translates and rotates in space in the way that is specified by the boundary conditions. Boundary conditions defined this way are linear. This leads to a form of BVP with in which \( r(x(t_0), x(t_f)) \) is defined as

\[
(2.6) \quad r(x(t_0), x(t_f)) = [x(t_0) - x_0; x(t_f) - x_f]
\]

where \( x_0 \) and \( x_f \) correspond to rectangular coordinates of the desired structures of the molecule at \( t = t_0 \), and \( t = t_f \), respectively.

From an analytical perspective the absolute boundary conditions might seem unnecessarily restrictive. In the study of conformation transitions, the change in relative location of atoms is of primary importance. Translational or rotational motion is expected to be irrelevant for most purposes. Boundary conditions may be defined so that the absolute locations of a molecule at the endpoints are not specified, but the conformation of the molecule does meet specific criteria that essentially determine the internal coordinates of the molecule. On each iteration, the internal coordinates of the desired final structure are projected (by optimal rotation and translation of the molecule) onto the location of the molecule at the end of the simulation. This leads to a form of BVP with nonlinear boundary conditions in which \( r(x(t_0), x(t_f)) \) is defined as

\[
(2.7) \quad r(x(t_0), x(t_f)) = [x(t_0) - ALIGN(m,x(t_0),x_0); x(t_f) - ALIGN(m,x(t_f),x_f)]
\]

To begin the explanation of the terms \( ALIGN(m,x(t_0),x_0) \) and \( ALIGN(m,x(t_f),x_f) \), let \( x_0 \) and \( x_f \) represent rectangular coordinates of the molecule in the conformations that are desired at the beginning and the end
of the conformational transition, respectively. The absolute location of $x_0$ and $x_f$ in the rectangular coordinate system is not important. Let $\zeta$ be an arbitrary $3n \times 1$ coordinate vector. We can define a function that transforms $\zeta$ into an $n \times 3$ representation of the coordinates. For notational convenience, let $\tilde{\zeta}$ represent this function applied to $\zeta$. Similarly, define $\eta(\tilde{\zeta})$ to be a function that transforms $\tilde{\zeta}$ from an $n \times 3$ representation of the coordinates into a $3n \times 1$ representation. Now, let $m$ be an $n \times 1$ vector of the atomic masses of the atoms of the system, let $\tilde{x}^{(m)}$ be an $n \times 3$ matrix with the mass-weighted mean of the conformation repeated in each row, and let $\mathcal{R}^{\text{min}}(m, \tilde{x}, \tilde{y})$ be the $3 \times 3$ rotation matrix that optimally rotates the conformation of an $n \times 3$ matrix, $\tilde{y}$, onto the conformation of an $n \times 3$ matrix, $\tilde{x}$ with respect to the mass vector, $m$. This optimal rotation matrix can be found using singular value decomposition as described in section 12.4 of [Gol1996]. Let the notation $||\cdot||_F$ is used to represent the Frobenius norm, and let $\hat{R}$ be an arbitrary rotation matrix. The Frobenius norm of an arbitrary matrix $A=\{a_{ij}\}$ can be computed using the formula $||A||_F=(\sum_i \sum_j |a_{ij}|^2)^{1/2}$. We may write

$$\mathcal{R}^{\text{min}}(m, \tilde{x}, \tilde{y}) = \arg \min_{\hat{R}} \|m^T (\tilde{x} - \tilde{y} \hat{R})\|_F^2,$$

A preferred method for finding $\mathcal{R}^{\text{min}}(w, \tilde{x}, \tilde{y})$ uses quaternions ([Cou2004]). In contrast to the method of [Gol1996], this method features a convenient way to exclude orthogonal transformation matrices that contain reflections. Now, define

$$\text{ALIGN}(m, x, y) = \eta(\tilde{x}^{(m)} + (\tilde{y} - \tilde{x}^{(m)}) \mathcal{R}^{\text{min}}(m, \tilde{x}^{(m)}, \tilde{y} - \tilde{x}^{(m)}))$$

So, the desired boundary conformations are optimally translated and rotated onto the locations of the molecule at $t = t_0$, and $t = t_f$ using the following formulas

$$\text{ALIGN}(m, x(t_0), x_0) = \eta\left(\tilde{x}^{(m)}(t_0) + (\tilde{x}_0 - \tilde{x}^{(m)}) \cdot \mathcal{R}^{\text{min}}(m, \tilde{x}(t_0) - \tilde{x}^{(m)}(t_0), \tilde{x}_0 - \tilde{x}_0^{(m)})\right),$$

$$\text{ALIGN}(m, x(t_f), x_f) = \eta\left(\tilde{x}^{(m)}(t_f) + (\tilde{x}_f - \tilde{x}^{(m)}) \cdot \mathcal{R}^{\text{min}}(m, \tilde{x}(t_f) - \tilde{x}^{(m)}(t_f), \tilde{x}_f - \tilde{x}_f^{(m)})\right)$$

Note that the method above would be functional for an arbitrary weight vector of non-negative real components. The choice of a weight vector of atomic masses is appealing since the center of mass of $\text{ALIGN}(m, x(t_0), x_0)$ and $\text{ALIGN}(m, x(t_f), x_f)$, have the same center of mass as $x_0$ and $x_f$, respectively. Finally,
note that a common measure of the similarity of conformation $x$ and conformation $y$ is given by the mass-weighted root mean squared deviation (RMSD) (after optimal rotation and translation) which can be computed as

$$\text{RMSD}(m,x,y) = \sqrt{\frac{\sum (x - \text{ALIGN}(m,x,y))^2}{n}}$$

(2.12)

2.2.2 Hamiltonians, equilibrium points, stability, and wells of a potential energy surface

To model conformational transitions by trajectories between local minima is convenient mathematically. It is noted, however, that, from a biological perspective, a conformation of a molecule may be classified as being in the conformation of a particular local minimum as long as it is approximately in that conformation. Both in reality and in the realistic theoretical application of Newtonian physics to an isolated system, the system will contain some energy and due to that energy, a molecule that is, from a practical perspective, in a certain conformation associated with a local minimum would not remain in precisely in that conformation, but would be subject to constant vibrational or oscillatory motion ‘around’ that local minimum. In fact, the precise conformation of the local minimum itself would be one of many possible conformations and may actually rarely be realized. This local motion may not affect the overall conformation from a qualitative or practical perspective. A more realistic model construction of boundary conditions might take into account this observation that molecules are in constant motion. Boundary conditions might be defined so that they would be satisfied as long as starting and ending conformations are appropriately ‘near’ the relevant local minima. More precision with respect to the usage of the terms ‘around’ and ‘near’ above will be provided later in this subsection. For now, we note that a conformation ‘around’ or ‘near’ a local minimum is a conformation that retains essential properties of the conformation of the local minimum. And, the set of all conformations satisfying this properties for a local minimum with name ‘$A$’ will be called the ‘$A$’ potential well, or ‘$A$’ well. A practical method for approximating a potential well will be presented in Chapter 4.
The momentum of a particle is equal to the product of its mass and velocity. A vector of momenta and a vector of coordinates in an arbitrary generalized coordinate system are commonly denoted by \( p \) and \( q \), respectively, with components \( p_i \) and \( q_i \) for positive integers \( i=1,\ldots,B \). The Hamiltonian or total energy function is a measurement of total energy. For an isolated system, using a rectangular coordinate system, the Hamiltonian is given by

\[
H(x,v,m) = E_{\text{TOTAL}}(x,m,v) = U(x(t)) + \frac{1}{2} \sum_{i=1}^{n} m_i \left[ \begin{array}{c}
\dot{y}_{3i-2}(t) \\
\dot{y}_{3i-1}(t) \\
\dot{y}_{3i}(t)
\end{array} \right]^2
\]

where \( m \) is an \( n \times 1 \) vector of atomic masses. Often the Hamiltonian is written in terms of generalized coordinates as

\[
H(p,q,M) = E_{\text{TOTAL}}(p,q) = U(q) + \frac{1}{2} \left\| M^{-1} p \right\|^2
\]

The equations of motion in terms of the Hamiltonian are given by

\[
\frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i}, \quad \frac{dq_i}{dt} = \frac{\partial H}{\partial p_i}, \quad i = 1,2,\ldots,B
\]

For a system of particles subject to a 2nd-order differential equation of the form of (2.1), the state of the system is by definition the position and momentum of all particles in the system. Phase space is a term that is used to refer to the set of all possible states of a system. Each possible state of the system corresponds to a unique point in the phase space. An equilibrium point, or critical point, or rest point, or fixed point, of an autonomous 2nd-order differential equation of the form of (2.1) is any point for which \( x'(t)=0 \) and \( x''(t)=0 \). Thus, if the coordinates of the system are such that the system is at a local minimum of then potential energy function, \( U(x) \) (i.e. \( M x''(t) = -\nabla U(x) = 0 \)), and all particles have zero velocity, (i.e. \( x'(t)=v(t)=0 \)), then the system is at an equilibrium point. The scope of this dissertation is limited to isolated systems for which \( H(x(t),m,v(t)) \) is constant with respect to time.

Stability of solutions is an important concept in the study of differential equations. Conceptually, we can think of stability of solutions as follows: If any two solutions are close to each other in phase space
at a point in time, \( t_0 \), then they will be close to each other for all \( t \geq t_0 \). A formal definition of stability is provided below followed by a related theorem:

(2.16) (definition) Let \( y^*(t) \) be a solution of a 1\(^{st}\)-order system of differential equations 
\[
y'(t) = h(t, y)
\]
then, \( y^*(t) \) is \textit{Lyapunov stable} on \( t \geq t_0 \) if, for any \( \varepsilon > 0 \), there exists a \( \delta(\varepsilon, t_0) > 0 \) such that
\[
\| y(t_0) - y^*(t_0) \| < \delta \implies \| y(t) - y^*(t) \| < \varepsilon
\]
for all \( t \geq t_0 \) where \( y(t) \) is any other solution of \( y'(t) = h(t, y) \)

([p. 221 of Jor1987]).

If \( \delta \) is independent of \( t_0 \), i.e. \( \delta(\varepsilon) \), then \( y^*(t) \) is \textit{uniformly stable} on \( t \geq t_0 \)

(2.17) (theorem) Let \( y^*(t) = 0, t \geq t_0 \), be the zero solution of a 1\(^{st}\)-order system of differential equations \( y'(t) = h(y) \). Then, \( y^*(t) \) is \textit{uniformly stable} on \( t \geq t_0 \) if there exists a \( V(y) \) with the following properties in some neighborhood \( N(y^*) \) of \( y = 0 \):

(i) \( V(y) \) and its partial derivatives are continuous;

(ii) \( V(y) \) is positive definite;

(iii) \( V'(y) \) is negative semidefinite.

Consider (2.17) where

(2.18) \[
y'(t) = h(y(t)), \quad t_0 < t < t_f
\]
is a 1\(^{st}\)-order system equivalent to (2.1). (See subsection 2.3.1 for an illustration of how an equivalent 1\(^{st}\)-order system can be created.) In fact, it will be possible to apply this theorem if the appropriate coordinate system is chosen. Suppose that \( x_0^* \) is an arbitrary isolated local minimum of \( U(x) \). For the coordinate system, let \( p \) be given by \( p = x - x_0^* \), and let \( p(t_0) = 0 \) which is equivalent to \( x(t_0) = x_0^* \). Assume that \( q(t_0) = 0 \). Then, \( y^*(t) = 0 \) is an equilibrium solution of (2.18). Based on these assumptions, it is shown on p. 283 of [Jor1987] that the Hamiltonian, \( H \), is a function satisfying the criteria of (2.17). So, \( y = 0 \) is a uniformly stable solution of (2.18).
There is a choice for the metric that will make the uniform stability result for Hamiltonian local minima particularly meaningful. First, note that if \( \text{ALIGN}(m,x,x') = 0 \) for any weight vector \( m \), then \( x \) and \( x' \) have the same internal coordinates and the same conformations. Any differences in rectangular coordinates are due to rotation and translation. Let’s define an equivalence class. For some weight vector \( m \),

\[
(2.19) \quad x: \{ x, x' \in \mathbb{R}^{3n} \cap \text{ALIGN}(m,x,x') = 0 \}
\]

By defining equivalence classes in this way for any set of coordinates, it was proven in [Ste2002] that a metric can be defined with respect to equivalence classes. Then, let us define

\[
(2.20) \quad \| x - y \|_{\text{RMSD}(m)} = \| x - \text{ALIGN}(m,x,y) \| / \sqrt{n}
\]

In this definition, \( x \) and \( y \) are \( 3n \times 1 \) vectors and can be considered to be representatives of their respective equivalence classes. The choice of the representative doesn’t affect the result.

Now, let’s consider an extended definition of the \( \text{RMSD}(m) \) metric. For any \( t \), let \( x(t) \) and \( v(t) \) be the position and velocity trajectories for an IVP of the form (2.1),(2.2). Let \( y(t) \) be a \( 6n \times 1 \) dimensional phase space function comprised of \( x(t) \) and \( v(t) \) for a solution to (2.1),(2.2). Using MATLAB syntax, one can write this as \( y(t) = [x(t); v(t)] \) since the ordering the components is arbitrary. Now, for \( y_a = [x_a; v_a] \) and \( y_b = [x_b; v_b] \), define \( \| y_a - y_b \|_{\text{RMSD}(m)} \) as follows:

\[
(2.21) \quad \| y_a - y_b \|_{\text{RMSD}(m)} = \text{RMSD}(w,x_a,x_b) + \left\| \left( \tilde{v}_a - \tilde{v}_a^{(m)} \right) - \left( \tilde{v}_b - \tilde{v}_b^{(m)} \right) \right\| \right\} \left\| x_a - x_a^{(m)} - x_b^{(m)} \right\|
\]

The first term on the right hand side is measure of the distance between \( x_b \) and \( x_a \) after \( x_b \) has been optimally translated and rotated onto \( x_a \). The last two terms on the right hand side collectively are a measure of the distance between \( v_b \) and \( v_a \) after \( v_b \) has been translated and rotated onto \( v_a \) using the rotation matrix for the rotation of \( x_b \) onto \( x_a \). Furthermore, let \( y_0 = [x_0; v_0] \) be a \( 6n \times 1 \) dimensional equilibrium point in phase space (comprised of position and velocity vectors). Then, \( y_0 \) is a stable equilibrium point if, given some \( t^* \) and any \( \varepsilon > 0 \), there is a \( \delta > 0 \) such that every solution for which the phase space function, \( y(t) \), satisfies

\[
(2.22) \quad \| y(t^*) - y_0 \|_{\text{RMSD}(m)} < \delta
\]
exists and satisfies
\[ (2.23) \quad \| y(t) - y_0 \|_{\text{RMSD}(m)} < \varepsilon \]
for all \( t \) satisfying \( 0 \leq t \leq t^* \). Recall that \( x_0^* \) is an isolated local minimum of \( U(x) \) and that the corresponding phase space equilibrium point \( y_0^* \) is stable. Note that if
\[ (2.24) \quad \| y(t) - y_0^* \|_{\text{RMSD}(m)} < \varepsilon \]
holds, then so does
\[ (2.25) \quad \| x(t) - x_0^* \|_{\text{RMSD}(m)} < \varepsilon \]
Moreover, if \( \delta \) and \( \varepsilon \) are chosen so that the conformation of a molecule retains properties (to be determined on a case-by-case basis) essential to the conformation of the local minimum from an analytical perspective, then we will say that the molecule remains in a well in the potential energy surface surrounding the local minimum \( x_0^* \). Informally, a well in a potential energy surface surrounding a local minimum is a region such that if the state of the system at some point in time is sufficiently close to the equilibrium point corresponding to the local minimum, the coordinates system will also remain measurably close to the local minimum and the kinetic energy of the system will always be bounded. Now, it should be pointed out that it remains to show that this extended definition of the \( \text{RMSD}(m) \) satisfies the properties of a metric. We will use the term \textit{potential energy well} to refer to an arbitrary well in a potential energy surface surrounding some local minimum. Later in this thesis, we will provide names for specific local minima of a specific molecule. The well surrounding a named local minima with name ‘A’ will be called the ‘A’ well, or ‘A’ potential well. For applications to be discussed in Chapter 4, it will be of interest to select a value for \( \varepsilon \) that is maximal in that it as large as it can be while conformations satisfying (2.25) still retain essential properties. Call this maximal value \( \varepsilon_{\text{max}} \). It is expected that empirical studies would be necessary to determine \( \varepsilon_{\text{max}} \) and a corresponding \( \delta_{\text{max}} \). Consider an IVP consisting of (2.1) and
\[ (2.26) \quad x(t_0) = x_0^*, \quad v(t_0) = v_0 \]
where \( x_0^* \) is an isolated local minimum of \( U(x) \) which is also stable. Below are some observations and assumptions related to this IVP:
1. If \( \|v_0\| < \delta_{\text{max}} \), then the conformation of a molecule will retain essential properties of the local minimum for the duration of the simulation. So, there will be no conformational transition.

2. If \( \|v_0\| > \delta_{\text{max}} \), then the conformation of a molecule may or may not retain essential properties of the local minimum from an analytical perspective for the duration of the simulation. So, there may or may not be a conformational transition. It is assumed that the likelihood of a conformation transition is directly correlated with the amount by which \( \|v_0\| \) exceeds \( \delta_{\text{max}} \) and for longer simulation times (i.e. large values of \( t_f - t_0 \)).

3. If \( \|v_0\| > \delta_{\text{max}} \), then whether or not there is a conformational transition and the type of conformational transition will be dependent on the direction of \( v_0 \).

4. A stable conformation can be expected to be associated with a local minimum of the potential energy function. The relative stability of a stable conformation and the likelihood of particular conformational transitions are influenced not only by the potential energy associated with that conformation but also by properties of the potential well (such as geometric shape) corresponding to that conformation.

2.3 Preliminaries II

2.3.1 Multiple shooting

A brief, informal description of the numerical method for solving BVP’s that is known as multiple shooting was given in section 1.2. For a more formal description of multiple shooting methods as they apply to AA-MDS, it is convenient to rewrite the system of \( 3n \) equations in the 2nd-order ODE (2.1) equivalently as a system of \( 6n \) equations that define a 1st-order ODE as shown in equation (2.27) below. Then, the BVP (2.1),(2.3) can be written in the form below

\[
\begin{align*}
(2.27) & \quad y'(t) = h(y(t)), \quad t_0 < t < t_f \\
(2.28) & \quad r(y(t_0), y(t_f)) = 0
\end{align*}
\]
where \( r(y(t_0), y(t_f)) \equiv r(x(t_0), v(t_0), x(t_f), v(t_f)) \). Since we are assuming \( R = 6n \) in this chapter, \( r(y(t_0), y(t_f)) \) is a function with \( 6n \) components.

### 2.3.1.1 Example 1: Transformation from 2nd-order to 1st-order BVP

To illustrate how this may be accomplished, consider the special case of two particles in three dimensions where the components of the vectors below have one or two subscripts with the first subscript corresponding to the particle number and, if present, the second subscript corresponds to rectangular coordinate number. Also, assume that the boundary conditions are given by a requirement on the absolute location in a rectangular coordinate system at each boundary point. That is, \( x(t_0) = x_0 \) and \( x(t_f) = x_f \). So, using MATLAB notation, we can write

\[
(2.29) \quad r(x(t_0), v(t_0), x(t_f), v(t_f)) = [x(t_0) - x^0; x(t_f) - x^f]
\]

Component-wise, let us use the notation

\[
(2.30) \quad x = [x_{11}; x_{12}; x_{13}; x_{21}; x_{22}; x_{23}],
\]

\[
(2.31) \quad v = [v_{11}; v_{12}; v_{13}; v_{21}; v_{22}; v_{23}],
\]

\[
(2.32) \quad a = [a_{11}; a_{12}; a_{13}; a_{21}; a_{22}; a_{23}],
\]

\[
(2.33) \quad f = [f_{11}; f_{12}; f_{13}; f_{21}; f_{22}; f_{23}],
\]

\[
(2.34) \quad x^0 = [x_{11}^0; x_{12}^0; x_{13}^0; x_{21}^0; x_{22}^0; x_{23}^0],
\]

\[
(2.35) \quad x^f = [x_{11}^f; x_{12}^f; x_{13}^f; x_{21}^f; x_{22}^f; x_{23}^f], \text{and}
\]

\[
M = \text{diag}([m_1; m_1; m_1; m_2; m_2; m_2])
\]

The last formula indicates that \( M \) is a \( 6 \times 6 \) diagonal matrix with the \( i \)th entry given by the \( i \)th component of the vector in the argument. To accomplish the transformation, we can write

\[
(2.36) \quad y = [x_{11}^i; v_{11}; x_{12}^i; v_{12}; x_{13}^i; v_{13}; x_{21}^i; v_{21}; x_{22}^i; v_{22}; x_{23}^i; v_{23}],
\]

\[
(2.37) \quad h = [v_{11}/m_1; v_{12}/m_1; v_{13}/m_1; v_{21}/m_2; v_{22}/m_2; v_{23}/m_2].
\]
If we let

\[
\begin{align*}
\mathbf{A}_1 &= \begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}, \\
\mathbf{A}_2 &= \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
\end{align*}
\]

then, we can write

\[
\begin{align*}
y &= \mathbf{A}_1 \mathbf{x} + \mathbf{A}_2 \mathbf{v} \\
h &= \mathbf{A}_1 \mathbf{v} + \mathbf{A}_2 (\mathbf{M}^{-1} \mathbf{f})
\end{align*}
\]

Furthermore, if we let

\[
\begin{align*}
\mathbf{B}_0 &= \begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
\end{align*}
\]

and
Then,

\[ r(y(t_0), y(t_f)) = B_0 y(t_0) + B_f y(t_f) - [x^0; x^f] \]

Then,

\[ r(y(t_0), y(t_f)) = B_0 y(t_0) + B_f y(t_f) - [x^0; x^f] \]

END OF EXAMPLE

For numerical solution of an ODE on an interval \([t_0, t_f]\), a mesh is a set of discrete points contained within the interval \([t_0, t_f]\). Continuing with the description of multiple shooting methods, following [Asc1995], we subdivide \([t_0, t_f]\) into \(N\) subintervals using a mesh \(\Delta_s : \{t_i : 0 \leq i \leq N\}\) such that \(t_0 < t_1 < \ldots < t_N = t_f\). Then, we will solve IVP’s on each subinterval:

\[ y'(t; s) = h(y(t; s)), \; y(t_i) = s_i, \; t_i < t < t_{i+1}, \; 0 \leq i \leq N-1 \]

where \(s = [s_0; s_1; \ldots; s_{N-1}]\) is a parameter vector in which each block component, \(s_i\), contains initial conditions at \(t = t_i\). The notation with the semicolon followed by \(s\) provides a reminder that the solution is dependent on the initial conditions which are determined by \(s\). There are \(6nN\) unknown parameters. The solution for a given \(s\) is

\[ y(t) \equiv y(t; s_i), \; t_i < t < t_{i+1}, \; 0 \leq i \leq N-1 \]

We want to find \(s^*\) such that
(2.40) \[ y_i(t_{i+1};s_i^*) = s_{i+1}^*, \quad 0 \leq i \leq N - 1 \]

(2.41) \[ r(s_1^*, y_N(t_f; s_{N-1}^*)) = 0. \]

We define

\[
F(s) \equiv \begin{bmatrix}
y_0(t_1; s_0) - s_1 \\
y_1(t_2; s_1) - s_2 \\
\vdots \\
y_{N-2}(t_{N-1}; s_{N-2}) - s_{N-1} \\
r(s_0, y_{N-1}(t_f; s_{N-1}))
\end{bmatrix},
\]

and want to find solutions of

(2.43) \[ F(s) = 0 \]

Various global convergence schemes can be used. For the applications to be described in this work, we have used two different iterative global convergence schemes for the above equation– (1) a dogleg trust region algorithm with residual reduction criterion and (2) a damped Newton algorithm with natural monotonicity reduction criterion. For any iterative approach with iterates given by \( s^0, s^1, \ldots, s^k \), we may write

(2.44) \[ s^{k+1} = s^k + \xi^k. \]

It is expected that, using either of the two global convergence schemes indicated above, \( \xi^k = \xi^{\text{Newton}(k)} \) in the final steps of a converging sequence where \( \xi^{\text{Newton}(k)} \) is the Newton step on the \( k \)th iteration. The Newton step is generated by solving

(2.45) \[ F'(s^k)\xi^{\text{Newton}(k)} = -F(s^k) \]

where \( F'(s) \) is the Jacobian of \( F(s) \). If \( F' \) is nonsingular, then

(2.46) \[ \xi^{\text{Newton}(k)} = -F'(s^k)^{-1} F(s^k). \]

Here,
\[
\begin{align*}
F'(s) &= \begin{bmatrix}
Y_0(t_1) & -I & 0 & \cdots & \cdots & 0 \\
0 & Y_1(t_2) & -I & \ddots & \ddots & \vdots \\
\vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\
\vdots & \ddots & \ddots & \ddots & \ddots & 0 \\
0 & \ddots & \ddots & 0 & Y_{N-2}(t_{N-1}) & -I \\
B_0 & 0 & \cdots & \cdots & 0 & B_f Y_{N-1}(t_f)
\end{bmatrix}
\end{align*}
\]

where

\[
Y_i(t) = \frac{\partial y_i(t; s_i)}{\partial s}, \quad 0 \leq i \leq N-1
\]

and

\[
B_0 = \frac{\partial}{\partial u} r(u, v), \quad B_f = \frac{\partial}{\partial v} r(u, v)
\]

at \(u=y(t_0; s)\) and \(v=y(t_f; s)\). For each \(i\), applying Theorem 7.1.8 of [Sto2002], we can find each \(Y_i\) numerically, step-by-step as we solve (2.40) by solving the following matrix ODE:

\[
\begin{align*}
\frac{d}{dt} Y_i(t; s_i) &= \frac{\partial}{\partial y} h(y_i(t; s_i)) Y(t; s_i), \quad t_i < t < t_{i+1} \\
Y_i(t_i) &= I, \quad 0 \leq i \leq N-1
\end{align*}
\]

Due to intrinsic error in the numerical methods for solving ODE’s and also because of numerical errors associated with finite arithmetic, the best that can be achieved is an approximate solution to the BVP. The accuracy of the approximation as a Newtonian trajectory and as a solution to the BVP can be expected to depend on the accuracy of the method for numerical solution of IVP’s, mesh selection for IVP solutions, MS mesh selection, and tolerances set for identifying a numerical solution to \(F(s)=0\).

### 2.3.1.2 Example 2: Determining the Jacobian

Continuing with the problem from Example 1, in this example, we wish to determine \(B_0\), \(B_f\) and \(\partial h/\partial y\). The boundary conditions are linear, so we can expect that the matrices \(B_0\) and \(B_f\) from (2.55) are constant. They are, in fact, the same matrices, \(B_0\) and \(B_f\) that were given above in Example 1. The Hessian matrix of the
potential energy function is a 6 x 6 matrix. Let the entry in the \( i \)th row and \( j \)th column of \( \nabla^2 U(x) \) be represented by \( \nabla^2 U_{ij} \). Then, \( \partial h / \partial y \) takes the form given below:

\[
\begin{bmatrix}
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\nabla^2 U_{11} & 0 & \nabla^2 U_{12} & 0 & \nabla^2 U_{13} & 0 & \nabla^2 U_{14} & 0 & \nabla^2 U_{15} & 0 & \nabla^2 U_{16} & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\nabla^2 U_{21} & 0 & \nabla^2 U_{22} & 0 & \nabla^2 U_{23} & 0 & \nabla^2 U_{24} & 0 & \nabla^2 U_{25} & 0 & \nabla^2 U_{26} & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
\nabla^2 U_{31} & 0 & \nabla^2 U_{32} & 0 & \nabla^2 U_{33} & 0 & \nabla^2 U_{34} & 0 & \nabla^2 U_{35} & 0 & \nabla^2 U_{36} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
\nabla^2 U_{41} & 0 & \nabla^2 U_{42} & 0 & \nabla^2 U_{43} & 0 & \nabla^2 U_{44} & 0 & \nabla^2 U_{45} & 0 & \nabla^2 U_{46} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
\nabla^2 U_{51} & 0 & \nabla^2 U_{52} & 0 & \nabla^2 U_{53} & 0 & \nabla^2 U_{54} & 0 & \nabla^2 U_{55} & 0 & \nabla^2 U_{56} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
\nabla^2 U_{61} & 0 & \nabla^2 U_{62} & 0 & \nabla^2 U_{63} & 0 & \nabla^2 U_{64} & 0 & \nabla^2 U_{65} & 0 & \nabla^2 U_{66} & 0
\end{bmatrix}
\]

(2.52)

In general, the \( Y_i \) matrices will be dense despite the structural sparseness of \( \partial h / \partial y \). END OF EXAMPLE

**2.3.1.3 Example 3: Application to Mueller potential**

Here is an example in which a single particle exists in two dimensions subject to the Mueller potential — a potential energy function that has been used as a case study for BV-AA-MDS algorithms (e.g. [Elb1999]). This example is presented to provide a simple illustration of the multiple shooting method. The BVP is of the form (2.1),(2.3). For planar rectangular coordinates \( x_1 \) and \( x_2 \), the Mueller potential is given by

\[
U(x(t)) = 200e^{(x_1-1)^2-10x_2^2} - 100e^{x_1^2-10(x_2-0.5)^2} - 170e^{-6.5(x_1+0.5)^2+11(x_1+0.5)(x_2-1.5)-6.5(x_2-1.5)^2} + 15e^{-0.7(x_1+1)^2+0.6(x_1+1)(x_2-1)-0.7(x_2-1)^2}
\]

(2.53)

The multiple shooting algorithm was applied for \( t_0=0 \) and \( t_f=2.9 \). Two local minima for the Mueller potential are \([x_1;x_2] \approx [-0.5582; 1.4420] \equiv x_0\) and \([x_1;x_2] \approx [0.6235; 0.0280] \equiv x_f\). The boundary value function is defined to be
\[(2.54) \quad r(x(t_0), v(t_0), x(t_f), v(t_f)) = [x(t_0) - x_0; x(t_f) - x_f].\]

A contour plot for the Mueller potential is shown in Figure 2.2. The dark blue areas represent low energy wells around the local minima. The multiple shooting algorithm was applied with four subintervals of uniform time. An approximate solution was found on the 5th iteration. Referring to Figure 2.2, the thick black line on this figure represents the position trajectory of the solution. The initial position trajectory is shown in light green. The position trajectories of the 2nd, 3rd and 4th iterations are in purple. Due to the similarity between the position trajectories of the 4th and 5th iterations, in the figure, the position trajectory of the 4th iteration is almost completely covered by the position trajectory for the 5th iteration.

END OF EXAMPLE

2.3.2 Single shooting

Multiple shooting applied with one subinterval, that is with \(N=1\), is often referred to as single shooting. In the case of single shooting, we have \(F(s)\)

\[(2.55) \quad F(s) = r(s_0, y_0(t_f; s_0)).\]

and the Jacobian, \(F'(s)\), takes the slightly different form

\[(2.56) \quad F'(s) = B_0 + B_f Y(t)\]

2.3.3 Finite difference methods

Finite difference methods are conceptually different than shooting methods. No IVP’s are solved. Rather than attempting to find parameters for a set of IVP’s that collectively solve the BVP (2.27),(2.28), one seeks to directly find a set of parameter values for each element of a mesh that satisfies the BVP on the mesh. In finite difference approaches, as will be described below, derivatives are replaced with difference quotients in the differential equation and boundary value condition equations. For nonlinear BVP’s, finite difference methods, like shooting methods, involve an iterative attempt to solve a system of nonlinear equations. Accuracy of the solution depends both on user-provided error tolerance settings, mesh selection, and accuracy of difference quotients described above. The content of this section is based on the
description of finite difference methods in Chapter 5 of [Asc1995]. The basic steps of most finite difference methods can be summarized as:

1. Set $k=0$ and choose a mesh, $\Delta^i$, consisting of $D+1$ points, $t_0, t_1, \ldots, t_D$, that satisfies

   $\Delta^i: t_0 < t_1 < \ldots < t_D = t_f$

2. Find a set of points, $x_A^k = [x_0^k, x_1^k, \ldots, x_D^k, x_D^k]$, representing the values of an initial position trajectory at the mesh points.

3. Replace derivatives with difference quotients in the differential equations and boundary conditions, and determine if $x_A^k$ solves the BVP to desired tolerances. If it does, then stop.

4. Otherwise, form a set of algebraic equations to modify the values of the position trajectory.

   This will result in another guess for a solution which we label as $x_A^{k+1}$.

5. Repeat steps 3 and 4 until convergence or until a maximum number of iterations is reached.

For a brief description of finite difference methods that is more detailed and more specific to AA-MDS, it is most convenient to describe the BVP in a modified form of (2.1), (2.3). We multiply the equation (2.1) by $M^{-1}$ and use the fact that $a(t)=x''(t)$ to arrive at the form

\[
(2.57) \quad x''(t) = M^{-1}f(x(t)), \quad t_0 < t < t_f
\]

\[
(2.58) \quad r(x(t_0), v(t_0), x(t_f), v(t_f)) = 0
\]

We will assume that the time step is constant, so that $h = t_{i+1} - t_i$ for all $i=0,\ldots, D-1$. We then can make a substitution in (2.57) based on the approximation

\[
(2.59) \quad x''(t) \approx \left( x(t_{i+1}) - 2x(t_i) + x(t_{i-1}) \right) / h^2
\]

to arrive at

\[
(2.60) \quad \left[ x(t_{i+1}) - 2x(t_i) + x(t_{i-1}) \right] / h^2 = M^{-1}f(x(t_i)) \quad 1 \leq i \leq D-1
\]

\[
(2.61) \quad r(x(t_0), v(t_0), x(t_f), v(t_f)) = 0
\]

where $h$ is appropriately small. (2.60) and (2.61) then form a system of algebraic nonlinear equations with $3n(D-1) + R = 3n(D-1) + 6n = 3n(D+1)$ equations. The parameters are the values of $x(t_i)$ for $0 \leq i \leq D$. So,
there are $3n(D+1)$ parameters. Different numerical methods for solving algebraic nonlinear equations can then be applied this system. As an example, we show how Newton’s method can be applied. As in the presentation in section 2.3.1, a vector function, $F(s)$, will be defined with nonlinear components. We seek a set of parameters, $s^*$, such that the vector equation $F(s^*) = 0$ is satisfied. The Newton step, $\xi^{\text{Newt}(k)}$, is the solution of

$$F'(s^k)\xi^{\text{Newt}(k)} := -F(s^k)$$

and the parameter update, $s^{k+1}$, is given by

$$s^{k+1} = s^k + \xi^{\text{Newt}(k)}.$$

For the multiple shooting method, there are a set of parameters and a trajectory associated with each iteration. As the method is described in section 2.3.1, the set of parameters is a subset of the trajectory. In finite difference method as described here, the set of parameters is the entire position trajectory. (And, the velocity trajectory can be computed from the position trajectory.) This equivalence is emphasized below in the redundant definition of the parameter set on the $k^{th}$ iteration, $s^k$, as $s^k \equiv x_{\Delta}^k$. We now describe application of Newton’s method for the finite difference scheme given above in more detail. For simplicity of presentation, let’s temporarily drop the superscript notation from above that indicates iteration number. Let

$$s \equiv x_{\Delta} \equiv [ x_0; \ldots ; x_N ] \equiv [ x(t_0); \ldots ; x(t_D) ]$$

represent a $3n(D+1)$ parameter vector and let $N_{\Delta}$ represent a difference operator on the mesh $\Delta'$ where

$$N_{\Delta}x_i = \frac{1}{h^2} [ x_{i+1} - 2x_i + x_{i-1} ] - M^{-1} \cdot f(x_i)$$

and let

$$F(s) \equiv [ N_{\Delta}x_1; \ldots ; N_{\Delta}x_{D-1}; r(x_0, x_D) ]$$

Here, to simplify notation, we assume that the boundary conditions are a function of $x_0$ and $x_D$ and are not a function of the initial velocities or terminal velocities. The Newton iteration (2.62) applied to (2.60), (2.61) for iteration $k+1$ results in

$$\frac{1}{h^2} [ \xi_{i+1} - 2\xi_i + \xi_{i-1} ] - M^{-1} A(x_i) \xi_i = -N_{\Delta}x_i$$

(2.67)

$$B_0 \xi_0 + B f \xi_D = r(x_0, x_D)$$

(2.68)
Here, the Newton step is $\xi^{\text{Newton}} = [\xi_0; \ldots; \xi_D]$ where each $\xi_i$ is a $3n \times 1$ vector. Note that $s^0 = x^0$ is an initial guess and $s^k = x^k$ for $k>0$ are known from the previous iteration. And,

\begin{align*}
(2.69) & \quad A(x_i) \equiv \frac{\partial^2 F}{\partial s^2}(x_i) = \nabla^2 U(x_i) \\
(2.70) & \quad B_0 = \frac{\partial}{\partial u}[r(u,v)], B_f = \frac{\partial}{\partial v}[r(u,v)] \text{ at } u = x_0, v = x_D
\end{align*}

The $(k+1)^{\text{st}}$ iterate is given by $x^{k+1} = x^k + \xi^{\text{Newton}(k)} \iff s^{k+1} = s^k + \xi^{\text{Newton}(k)}$.

### 2.3.4 Global convergence approaches

The shooting methods and finite difference methods described above are iterative methods for solving nonlinear BVP’s. For a given iteration, $k$, the nonlinear BVP is discretized and the numerical method leads to the calculation of $F(s^k)$, where $F$ is a vector-valued function $F$, of the parameter vector $s^k$. In the descriptions above, Newton’s method for nonlinear equations is used as a way to attempt to converge toward a parameter vector, $s^*$ that satisfies $F(s^*) = 0$. The value of $F$ for a given parameter vector is often called the residual, or residual vector. In practice, a solution is found when $s^k \approx s^*$ so that $F(s^k) \approx 0$.

Local convergence of Newton’s method for nonlinear equations to a solution can be guaranteed in theory under certain conditions. The conditions tend to be difficult to verify in practice. A practical implementation strategy for Newton’s method with nonlinear equations is to use a problem-specific strategy or a systematic strategy to generate a set of trial initial parameter vectors. Trial initial parameters vectors that result in initial trajectories with smaller residuals would be preferred initial parameter vectors. The importance of the initial guess in determining eventual convergence can be reduced by using a globally convergent modification of Newton’s method. Such modifications are described in section 6.5.3 of [Den1996], Chapter 8 of [Asc1995], and Chapter 8 of [Deu2002]. Other methods for solving nonlinear equations could also be employed.

Two globally convergent modifications of Newton’s method for nonlinear equations are described and applied in this dissertation. It could be argued that Newton or Newton-like methods are, in general, rather expensive computationally and not appropriate for BV-AA-MDS with a sizable number of atoms.
This argument should be addressed, so some explanations for the use of globally convergent modifications of Newton’s method for BV-AA-MDS are provided below:

1. In the general application of globally convergent modifications of Newton’s method, analytic methods for computing $F'(s)$ are often not available. Computing $F'(s)$ by finite differences can be computationally expensive as it can be expected to require an additional IVP solve for each component of $s$. In the case of BV-AA-MDS, the Hessian matrix of a biomolecular potential function, $U(x)$, can be evaluated analytically, and this efficient Hessian matrix evaluation can be used in the calculation of $F'(s)$ by solution of the variational equation as described in section 2.3.1. While this approach also requires additional IVP solve’s, they can be performed step-by-step when solving the initial value problems that determine $F(s)$ and do not require additional force field evaluations.

2. For the case where the number of multiple shooting parameters and the number of boundary conditions are both equal to $6n$, $F'(s)$ is $6n \times 6n$. For large $n$, solving the linear system $F'(s)\xi = F(s)$ can be computationally expensive. Because of the almost block diagonal structure of $F'(s)$, however, this linear system solve is $O(N(6n)^3)$ rather than $O((6nN)^3)$ ([Deu2004]).

3. The high cost of solving $F'(s)\xi = F(s)$ can be reduced by reducing the number of parameters from $6nN$ to a significantly smaller number. Also, the number of boundary conditions may also possibly be reduced significantly from $6n$ to a significantly smaller number. For a 22-atom alanine dipeptide, useful results were obtained for single shooting with 5 parameters instead of $6 \times 22 = 132$ and 50 boundary conditions instead of 132. Methods for solving $F'(s)\xi = F(s)$ when the number of boundary conditions and number of parameters may be reduced will be addressed in Chapter 4.

4. For a case study where $n=253$, the author has been able to solve $F'(s)\xi = F(s)$ in about one second of real time using MATLAB. There are applications with $n=253$ of scientific and computational interest which would be appropriate for use in the study of the BV-AA-MDS multiple shooting approach.
2.3.5 Identifying a numerical solution

As mentioned above, to determine whether one has solved the problem \( F(s) = 0 \) numerically, one can make an assessment of the distance between \( F(s) \) and the zero vector. A reasonable strategy could be to consider the parameter vector, \( s \), to be a parameter vector that solves the BVP when \( s \) satisfies
\[ \| F(s) \|_2 < \text{TOL}_F \] where \( \text{TOL}_F \) is a tolerance indicating the largest acceptable value of \( \| F(s) \|_2 \) for a solution trajectory. It is sometime useful to scale the components of \( F(s) \). For the purpose of scaling, the vector \( F \) can be left multiplied by a \( 3n \)-dimensional diagonal scaling matrix, \( W \), so that the solution identification criterion is \( \| WF(s) \|_2^2 < \text{TOL}_{WF} \). For example, in some BV-AA-MDS multiple shooting algorithms that are implemented in this work, the assignment \( W = M^{1/2} \) is used, resulting in scaling based on atomic mass of the atoms. The solution identification criterion \( \| F(s) \|_2 < \text{TOL}_F \) can be considered to be a special case of the more general stopping criterion in which the diagonal scaling matrix is the identity matrix. The intuition behind the atomic mass scaling is an argument that deviation of a heavy atom from a desired location should be given greater weight than a similar deviation for a hydrogen atom, since hydrogen atoms may exhibit local vibrational motions at a higher frequency and with larger magnitudes.

2.3.6 Globally convergent modifications of Newton’s method

Below two different globally convergent modifications of Newton’s method are briefly described. Using the notation of section 2.3, the step is the adjustment, \( \xi^k \), to the parameter vector, \( s^k \), for the next iteration. In each approach there is a criterion to determine whether or not the Newton step will likely result in a long term progress toward a solution. The likelihood of long term progress is based on some measure of incremental progress toward a solution. If the Newton step is not chosen, another step of smaller magnitude is chosen. In the first modification, the step of smaller magnitude will also have a different direction. In the second modification, the direction for the step of smaller magnitude is the Newton direction.
2.3.6.1 Trust-region model with dogleg step

Trust region models are described in detail in section 6.5 of [Den1996] and Chapter 4 of [Noc2002]. Within the realm of trust region models, there are different approaches for proposing and accepting the next step. One approach is the dogleg approach. Below is a brief description of the trust region model with a dogleg step, adapted from the presentations in [Den1996] and [Noc2002], that was implemented for BV-AA-MDS using MS.

We seek $s$ such that $F(s) = 0$. A reasonable global convergence strategy is to try to find $s$ to solve

$$\min_s \Phi(s) \quad \text{where} \quad \Phi(s) = \frac{1}{2} F(s)^T F(s) = \frac{1}{2} \|F(s)\|^2.$$  

An approach for defining a reduction criterion in terms of $\Phi(s)$ will be now be described. Denoting the $i^{th}$ component of $F$ by $F_i$, it can be shown that

$$\nabla \Phi(s) = \frac{d}{dx} \sum_i \frac{1}{2} (F_i(s))^2 = F'(s)^T F(s)$$

A quadratic approximation for $\Phi(s + \xi)$ is given by $\hat{\Phi}(s, \xi)$ where

$$\hat{\Phi}(s, \xi) = \Phi(s) + \nabla \Phi(s)^T \xi + \frac{1}{2} \xi^T \nabla^2 \Phi(s) \xi.$$  

The Hessian can be written as

$$\nabla \Phi^2(s) = F'(s)^T F'(s) + \sum_i F_i(s) \nabla F_i^2(s).$$

If the Hessian is approximated by $F'(s)^T F'(s)$, then we have the approximation, $\tilde{\Phi}(s, \xi)$, for $\Phi(s + \xi)$ where $\tilde{\Phi}(s, \xi)$ is given by

$$\tilde{\Phi}(s, \xi) = \Phi(s) + \nabla \Phi(s)^T \xi + \frac{1}{2} \xi^T F'(s)^T F'(s) \xi.$$  

The Newton step for the system of NLE’s given by $F(s) = 0$ is

$$\xi^{\text{Newton}} = -F'(s)^{-1} F(s).$$
It can be shown that $\xi = \xi^{\text{Nwtn}}$ is a descent direction for the quadratic model defined by (2.75). This is a direct result of the fact that the Newton step for the quadratic model defined by (2.75), $\xi^{\Phi}$, is equal to $\xi^{\text{Nwtn}}$ as is shown here:

$$
(2.77) \quad \xi^{\Phi} = -(F'(s)^T F'(s))^{-1} \cdot \nabla \Phi(s) = - (F'(s)^T F'(s))^{-1} F'(s)^T F(s) = - F'(s)^{-1} F'(s)^T F(s) = - F'(s)^{-1} F(s) = \xi^{\text{Nwtn}}.
$$

The Cauchy step, $\xi^{\text{Cchy}}$, is the name for a step chosen in the opposite direction of the gradient with a magnitude chosen to minimize the quadratic model in the direction of $\nabla \Phi(s)$. It is

$$
(2.78) \quad \xi^{\text{Cchy}} = -\lambda \nabla \Phi(s) \text{ where } \lambda = \frac{\|\nabla \Phi(s)\|^2}{\nabla \Phi(s)^T F'(s)^T F(s)}
$$

where $\lambda$ is a positive scalar that determines the magnitude of the Cauchy step. It can be shown that $\|\xi^{\text{Cchy}}\| \leq \|\xi^{\text{Nwtn}}\|$. The dogleg step produces an approximate solution to the trust-region problem

$$
(2.79) \quad \min_{\xi} \bar{\Phi}(s + \xi) \text{ subject to } \|\xi\|_2 \leq \delta.
$$

Within the adaptive algorithm, the value of $\delta$ can vary from one iteration to the next. An algorithm that describes the step choice follows:

$$
(2.80) \text{ (algorithm): If } \|\xi^{\text{Nwtn}}\|_2 < \delta, \text{ choose } \xi = \xi^{\text{Nwtn}}.
$$

else, if $\|\xi^{\text{Nwtn}}\|_2 > \delta$ and $\|\xi^{\text{Cchy}}\|_2 \geq \delta$, choose $\xi = \delta \xi^{\text{Cchy}} / \|\xi^{\text{Cchy}}\|_2$.

else, if $\|\xi^{\text{Nwtn}}\|_2 > \delta$ and $\|\xi^{\text{Cchy}}\|_2 < \delta$, choose $\xi = \xi^{\text{Cchy}} + \nu (\xi^{\text{Nwtn}} - \xi^{\text{Cchy}})$

$$
\text{ where } \nu \text{ is the positive root of the equation } \|\xi^{\text{Cchy}} + \nu (\xi^{\text{Nwtn}} - \xi^{\text{Cchy}})\|_2^2 = \delta^2.
$$

So, the dogleg step will be either the Cauchy step, the Newton step, or a linear combination of these two steps. The sufficient reduction criterion is that

$$
(2.81) \quad \Phi(s + \xi) \leq \Phi(s) + \alpha [\nabla \Phi(s)]^T \xi
$$
for some small positive constant α. An adaptive approach for modification of δ, suggested and described in [Den1996], has been implemented. Stopping criteria are suggested and described in Chapter 7 of [Den1996].

An analogous approach is used with the global convergence strategy

\[ \min_{s \in \mathbb{R}^n} \Phi(s) = \frac{1}{2} \|WF(s)\|^2 \]

using the fact that for \( Z(s) = WF(s) \), we have \( Z(s) = WF(s) \). For more information on these global optimization methods, see Chapter 5 of [Den1996] and/or Chapter 3 of [Noc2002].

**2.3.6.2 Damped Newton model with the natural criterion function**

A damped Newton step is a step chosen in the direction of the Newton step with a magnitude may be smaller than the magnitude of the Newton step. The magnitude is adaptively chosen to satisfy a sufficient reduction criterion. A reduction criterion function that is commonly used is conjunction with a damped Newton step is the **natural criterion function** which results from considering an adaptive scaling matrix for each iteration defined by \( W = F'(s_k)^{-1} \) for the reduction criterion function. The scheme resulting from using a damped Newton step with the natural criterion function has some desirable properties and has been of practical use in applications for solving BVP’s using multiple shooting methods; it is described in [Deu2004], [Deu2002], and in Chapter 8 of [Asc1995]. The following description is adapted from the presentation in Chapter 8 of [Asc1995].

If \( s \) is the current parameter vector, \( \xi_{\text{Newton}} \) is the parameter adjustment computed by the Newton step, and \( \lambda \) is the damping factor, the natural criterion function, \( \Phi \), is

\[ \Phi(s + \lambda \xi_{\text{Newton}}) = \frac{1}{2} \|WF(s + \lambda \xi_{\text{Newton}})\|^2 = \frac{1}{2} \|F'(s)^{-1}F(s + \lambda \xi_{\text{Newton}})\|^2. \]

Note that for \( \lambda = 0 \),

\[ \Phi(s) = \frac{1}{2} \|F'(s)^{-1}F(s)\|^2 = \frac{1}{2} \|\xi_{\text{Newton}}\|^2. \]

Also for \( \lambda = 0 \), note that \( \nabla \Phi(s) \) is computed regarding \( F'(s)^{-1} \) as fixed. So, substituting into the result of (2.72), we have

\[ \nabla \Phi(s) = \left( F'(s)^{-1} F'(s) \right)^T F'(s)^{-1} F(s) = F'(s)^{-1} F(s) = -\xi_{\text{Newton}}. \]
The Newton direction is a descent direction for $\Phi(s + \xi Nwtn)$ since
\begin{equation}
\left( \xi Nwtn \right)^T \nabla \Phi(s) = \left( \xi Nwtn \right)^T \left( -\xi Nwtn \right) = -\| \xi Nwtn \|^2 = -2\Phi(s) < 0 .
\end{equation}

The damped Newton method is a backtracking line search method for the scalar $\lambda$ that satisfies
\begin{equation}
\Phi(s + \lambda \xi Nwtn) \leq \Phi(s) + \alpha \lambda \left[ \nabla \Phi(s) \right]^T \xi Nwtn
\end{equation}
for some small positive constant $\alpha$. If we include subscript notation to represent iteration $k$, make a substitution using (2.86) in (2.87), and define
\begin{equation}
\xi^{k+1} = -F'(s^k)^{-1} F(s^{k+1}) = -F'(s^k)^{-1} F(s + \lambda \xi Nwtn(k)) ,
\end{equation}
then this criterion can be rewritten as
\begin{equation}
\left\| \xi^{k+1} \right\|^2 \leq (1 - 2\alpha\lambda) \left\| \xi Nwtn(k) \right\|^2 .
\end{equation}

When we find the scalar $\lambda$ that satisfies this criterion, we set
\begin{equation}
\xi^k = \lambda \xi Nwtn \text{ and } s^{k+1} = s^k + \xi^k
\end{equation}

Suggested stopping criteria and approaches for predicting and modifying $\lambda$ are given in section 8.1 of [Asc1995]. The damped Newton model with the natural criterion function has been a popular global convergence scheme for numerical solution of BVP’s using MS. A further description of this application can be found in [Asc1995], [Deu2002], and [Deu2004]. Qualitatively, this optimization method with the natural criterion function has a reputation for facilitating rapid convergence when used in MS applications by permitting larger steps than with other reduction criteria. And, in the event, that the algorithm does not converge, the method has a reputation for stopping quickly and avoiding wasting time by taking tiny steps while waiting for a convergence that is not likely to occur.

### 2.3.7 Stochastic difference equation methods

The multiple shooting method and the finite difference method were not developed specifically for application to BV-AA-MDS. They are two common general numerical methods for solution of BVP’s for ODE’s. On the other hand, the stochastic difference equation (SDE) approach, introduced in [Ole1996]
and [Elb1999], was developed specifically for application to BV-AA-MDS. Physical motivation and physical significance was emphasized in the presentation of this method. In this section, it is asserted that, as a net result, the SDE approach is equivalent to a particular finite difference method.

The stochastic difference equation method developed in [Ole1996] is a method for the BVP given by (2.1), (2.3) when the beginning coordinates, \( x_0 \), and ending coordinates, \( x_f \), are known. So, the BVP takes the form

\[
\begin{align*}
M a(t) &= f(x(t)), \quad t_0 < t < t_f \\
r(x(t_0), v(t_0), x(t_f), v(t_f)) &= [x(t_0) - x_0 ; x(t_f) - x_f] = 0
\end{align*}
\]

The starting point for this method is the principle of least-action. According to the classical principle of least-action, the equations of motion are satisfied by finding a stationary point of the classical action

\[
S_{cl}(x(t)) = \int_{t_0}^{t_f} M \| v(t) \|^2 - U(x(t)) dt
\]

It is asserted in [Ole1996]) that upon differentiation, setting the result equal to 0, and discretization assuming a constant time step of \( \Delta t \), the classical action leads an optimization problem with undesirable characteristics. As an alternative, the Onsager-Machlup action is considered. This action is given as

\[
S_{OM}(x(t)) = \int_0^{t_f} \| Ma(t) + f(x(t)) \|^2 dt
\]

Upon differentiation, setting the result equal to zero, and discretization, the Onsager-Machlup action leads to the optimization problem

\[
\min_x \sum_{k=1}^{n} \left[ \frac{M}{\Delta t^2} \left[ x(t) + x(t - 2\Delta t) - 2x(t - \Delta t) \right] + f(x(t - \Delta t)) \right]^2
\]

subject to \( x(t_0) = x_0 \) and \( x(t_f) = x_f \)

One can obtain the identical optimization problem by defining an objective function analogous to that of equation (2.71) for the system of nonlinear equations given by (2.60), (2.61). As a net result, then, the method of [Ole1996] is equivalent to a finite difference approach with a global convergence optimization scheme.
In [Elb1999], the optimization problem of equation (2.95) is replaced by

\[
(2.96) \quad \min_{\Delta} \sum_{k=1}^{n} \left( \frac{M}{\Delta^2} [x(t) + x(t - 2\Delta t) - 2x(t - \Delta t)] + f(x(t - 2\Delta t)) \right)^2
\]

subject to \( x(t_0) = x_0 \) and \( x(t_f) = x_f \)

The subtle change, replacement of \( f(x(t - \Delta t)) \) with \( f(x(t - 2\Delta t)) \), reduces the local accuracy of the discretization method from \( O(\Delta t^4) \) to \( O(\Delta t^3) \). (For a definition of local accuracy and an explanation of the \( O(\cdot) \) notation in this context, see the appendix of this chapter, section 2.6.) This modification is considered because it results in a simplified calculation of the Jacobian, \( F'(s) \). (Note: The previously mentioned SDE method can also be applied with length (SDEL), rather than time (SDET) as the independent variable. As a numerical method, it is argued in [Elb2003b] that SDEL has some advantages in comparison to SDET. A drawback of SDEL is that if only approximate trajectories are obtained, then, the time scale of the trajectories can only be approximated. See also [Elb2002] for more information about SDEL.)

A feature of SDE that is emphasized in the publications of Ron Elber, e.g. [Elb2002], [Ole1996], is the potential for application with a step size that is too large to recover the exact trajectory. The argument is made that the resulting trajectory is a good approximation to an exact trajectory. In fact, it is argued that application of the SDE methods typically results in an automatic filtering of the high-frequency local motions of an exact solution. And, in contrast, the low-frequency components of an exact solution tend to be automatically retained.

\subsection*{2.3.8 Background information for biomolecular simulation}

\subsubsection*{2.3.8.1 Classical mechanics and quantum mechanics}

In theory, the most accurate models of molecular motion would be based on principles of quantum mechanics. AA-MDS as it has been described in the dissertation is an application of classical mechanics. While the usefulness of models based on classical mechanics, in general, is limited, if a system satisfies certain conditions, then models based on classical mechanics are believed to useful approximations. This dissertation uses a twenty-two atom system as a primary case study. This system could be modeled using a
quantum mechanical model or a hybrid model—one that incorporates some elements of quantum mechanics and some elements of classical mechanics. A primary motivation for choosing this case study is for its relevance as a model for larger macromolecules. For macromolecules of size greater than 100 atoms, quantum mechanical simulation is considered to be prohibitive due to computational cost ([Swa2005]). Moreover, for macromolecules, particularly at a temperature near that of their normal biological environment, the classical mechanical approximation is considered to be sufficiently accurate ([Erc1997],[Ste2003]). This dissertation will not consider simulation methods that incorporate quantum mechanics. Brief discussion of quantum mechanical methods can be found in [Erc1997], and [Ste2003]. More detailed discussion can be found in [Sch2002], [Lea2001], and [Fre2002].

2.3.8.2 Characterization of protein conformations

A protein produced via translation from mRNA, initially exists in an unfolded state. In ‘Thermodynamics of Protein Folding and Stability’ by Alan Cooper, (Chapter 6 of Protein: A Comprehensive Treatise, [Coo1999]), an unfolded state is characterized as

an ill-defined state, or rather set of states comprising anything that is not recognisably folded. A population of conformations, spanning and sampling wide ranges of conformation space depending on conditions. Usually quite open, irregular, heterogeneous, flexible, dynamic structures - no one molecule is like another, nor like itself from one moment to another...

For a protein to carry out its biological function, it must transition from the unfolded state to the folded state. This process is called folding. The folded state is characterized in [Coo1999] as

the biologically active (“native”) form of the [protein] (usually). Compact, showing extensive average conformational homogeneity with recognisable regions of regular, irregular and motif structures, on a background of dynamic thermal fluctuations...

Important measurements for assessing the state of a protein are the values of the pairs of dihedral angles around the C\(^\alpha\) atoms of the amino acids that make up the peptide chain. These angles are commonly called
φ and ψ angles. For an arbitrary amino acid, φ and ψ are defined in a way that is analogous to the way that φ and ψ are defined for the alanine amino acid of N-acetyl-N′-methylalaninamide in section 1.3. Pairwise interatomic distances are also important measurements for assessing the state of a protein. In the unfolded state, the distribution of φ and ψ angles can be quite broad (but not random). By contrast, in a folded state, the φ and ψ angles of many residues will typically populate a specific and narrower range of values.

The characterization in the excerpt from [Coo1999] clearly and accurately gives the impression that the folded state is not a single conformation, but, rather, an collection of conformations with some identifiable features in common. The folded state is a relatively stable state. And, often, the folded state is essentially homogenous as is suggested in this excerpt. But, the stability and homogeneity of the folded state can vary for different proteins. For some proteins, there may be alternate folded states. Some alternate folded states may be detrimental to the cell. In these cases, the alternate folded states may be called a misfolded state. For some proteins, there may also be intermediate states that exhibit some stability. These states are known as metastable states. There always exist equilibria among different folded states, metastable states, unfolded states, and misfolded states. The equilibria can be affected by environmental changes. In many cases, the conformational transitions of molecules between a folded state and another state (e.g. important metastable states, alternate folded states, or misfolded states) can have a vital bearing on important biological processes. The importance of conformational transition of molecules is a motivating factor for the study of BV-AA-MDS and methods for numerical solution for BV-AA-MDS.

IV-AA-MDS also has important biological applications. As an example, IV-AA-MDS may be applied to study local motions near a stable conformation of a biomolecule. In this case, initial values may include a stable structure and randomly assigned momenta scaled to satisfy some desired level of total energy in the system. IV-AA-MDS has also been applied to the study of folding and conformational transitions ([Kim2003], [Zag2001], [Rhe2003]). Loosely defined but generally accepted theories of protein folding and conformational transitions have emerged (e.g. [Onu2004]) and some of the collective understanding of these processes can be attributed to IV-AA-MDS. But, many questions about the folding
and transitional processes remain unanswered and attempts at verifying the existing theories are often anecdotal or incomplete.

### 2.3.8.3 System of units

Unless otherwise indicated, the AKMA system of units will be used in this dissertation. This is the system of units used in the MDS software CHARMM. AKMA is an acronym representing units of distance, energy, and mass. In the AKMA system of units, Angstroms (Å), Kilocalories/Mole (kcal mol\(^{-1}\)), Atomic mass units are the units of measure for distance, energy, and mass, respectively. Using this system, one AKMA unit of time is approximately \(4.888821 \times 10^{-14}\) seconds. Angles will always be given in degrees.

### 2.3.9 Computational limitations of AA-MDS and implications

IV-AA-MDS has some well-known limitations. The relevance of an AA-MDS trajectory is partially limited by the accuracy of the previously described potential energy function. For relevant discussion of issues related to the potential energy function, see [Gar2003], [Pri2002], [Go1983], [Gar2002], [Ued1978], [Ren2006], [Roy2005], [Hu2003]. For our purpose, we do not address this issue, but assume that the potential energy function is adequate for modeling purposes. A concept of statistical mechanics that is important to consider when analyzing trajectories is that the larger a sample of trajectories is, the more useful it may be for making inferences about the ensemble or population of trajectories to which it belongs.

From the perspective of the author, the most important limitations with respect to simulation of folding or conformational transitions of large molecules are computational in nature. The extent to which these limitations are realized depends on the number of molecules being simulated, the length of expected waiting times for the events to initiate, the expected duration of the events once they have been initiated. The former limitation will be referred to as the **system size limitation** and the remaining two will be grouped together and called the **time interval limitation**. At first glance, it might seem reasonable to consider AA-MDS for the collective activities of a simple one-celled organism. However, even a simple
one-celled organism is estimated to have hundreds of trillions of atoms. The system size limitation is a realization that the scope of a molecular dynamics simulation must be limited.

Computing an approximate solution to an IVP numerically involves a sequential iterative process of determining a discrete evolution of the system of particles in time on a mesh, or ordered set of time points. To capture the fastest motions which have vibrational frequencies on the order of $10^{-13} \text{s}$, the time steps, or increments between adjacent points of the mesh, are require to be about $10^{-15} \text{s}$. Because of this constraint on the length of the time steps, there are practical limits for the total time interval for a simulation, which are currently on the order of nanoseconds ($1 \text{ns} = 10^{-9} \text{s}$) or microseconds ($1 \text{μs} = 10^{-6} \text{s}$), where the range is due to factors such as size of the system (i.e. numbers of particles) to be simulated, computational resources available, and sample size required. In nature, conformational transitions of biological interest can occur, however, over much longer time intervals (e.g. milliseconds ($1 \text{ms} = 10^{-3} \text{s}$) or even seconds. This is the essence of the time interval limitation. Moreover, a single trajectory or small sample of trajectories may be useful, but, in the use of molecular dynamics simulation to study conformational transitions, it is often desirable to be able to make probabilistic or statistical assessments about characteristics a large population of transitional trajectories. So, a relatively large ensemble of transitional trajectories may be necessary to facilitate reliable inference about a population of transitional trajectories. This need exacerbates the chasm between desired and available computational resources. In light of the important limitations, how are and how can interesting, but computationally challenging, folding and transitional events be studied?

With respect to this dissertation, given the important computational limitations of AA-MDS, a series of three questions naturally arises from the perspective of the author. First, why should AA-MDS even be considered as a means for studying conformational transitions of biological interest? Second, if it can be justified, what are the advantages and disadvantages of engaging in this study with a boundary value approach rather than the more conventional initial value approach to AA-MDS? And, third, if a
BV-AA-MDS can be justified, why is the multiple shooting method worthy of consideration as a numerical method? These questions are addressed in the next section (section 2.4).

2.3.10 Distance matrix interpolation

Given a metric, or measure of distance, such as the Euclidean metric, for any conformation of a system of \( n \) particles, a distance matrix can be created. A distance matrix is simply a matrix with a metric between particle \( i \) and particle \( j \) stored as an entry in row \( i \), column \( j \). The distance matrix is unique. Because of elementary properties of a metric (e.g. [Ste2002]), it is also symmetric with zeros in all of the diagonal entries. Conversely, given a complete distance matrix, one can construct a set of coordinates. The set of coordinates is not unique. For a conformational transition of a system from one conformation to another with discrete snapshots for a given number of intermediate structures, a distance matrix can be constructed for each intermediate structure. In Chapter 5, methods are described and introduced to generate a sequence of distance matrices if the actual distance matrices for a Newtonian conformational transitions is not known, but only the distance matrices for beginning and ending conformations, and, perhaps, some intermediate conformations are available. From these distance matrices and with added constraints to uniquely determine the set of coordinates in a way that preserves spatial continuity, a meaningful sequence of conformations can be constructed. In this dissertation, a sequence constructed this way will be called a DMI position trajectory. As will be described in subsection 3.4.3, a trajectory generated by an approximate method such as a DMI can be used to generate initial trajectories that are necessary for application of numerical methods like MS to BV-AA-MDS. The following examples are intended to provide an intuitive understanding of two different types of DMI, linear DMI and nonlinear DMI. Linear DMI is DMI in which the evolution of all the interatomic distances proceeds monotonically with uniform relative rate. According to this definition, it is possible that the evolution might not proceed linearly with respect to time. However, what the definition does stipulate is that at any point in time, that all interatomic distances have proceeded in identical proportions from the beginning interatomic distance to the ending interatomic distance. Nonlinear DMI is DMI which is not linear as defined above. In the DMI approach, each snapshot of the
position trajectory is thought to represent an approximate MDS trajectory at some point in time between the initial time, $t_0$, and the ending time, $t_f$. However, the exact time in not determined by the model. We will call a trajectory with this property a hidden time trajectory. One can still construct a velocity trajectory by creating a DMI position trajectory with a dense mesh and use a discrete approximation to compute approximate velocities. We can expect that even if the DMI position trajectories are realistic the velocity trajectories may not be consistently realistic.

2.3.10.1 Example 4: Linear DMI for transformation of a triangle

Here we consider the transformation of a $3 \times 4 \times 4$ isosceles triangle to a $0.75 \times 5.75 \times 5.75$ isosceles triangle by linear DMI. In Figure 2.7, a transformation satisfying linear DMI is depicted. Also, in Figure 2.7 the interatomic distances are plotted sequentially. Note that all three sets form straight lines.

END OF EXAMPLE

2.3.10.2 Example 5: Nonlinear DMI for transformation of a triangle

Here we consider the transformation of a $3 \times 4 \times 4$ isosceles triangle to a $0.75 \times 5.75 \times 5.75$ isosceles triangle by an unspecified nonlinear DMI method. In Figure 2.8, a transformation based on a possible nonlinear DMI method is depicted and the interatomic distances are plotted sequentially. Note that the three sets when plotted all form different nonlinear patterns. END OF EXAMPLE

2.4 Ideas, Methods, and Analysis II

2.4.1 Why AA-MDS?

The ability to study of molecular processes experimentally is continually progressing due to technological and methodological advances. It is desirable to assess, at least at some level of detail, the accuracy and validity of inferences from AA-MDS by direct comparison with experimental data. If this option is not available, results could be questioned because they are derived from a simulation method with many assumptions that are difficult to verify. Regardless, in many cases, experimental methods still do not
allow the microscopic and dynamic view offered by AA-MDS, and, despite some shortcomings, the analysis and inferences resulting from AA-MDS are often given substantial weight to the study of molecular processes. How can a molecular dynamics problem be studied by a simulation approach when a direct AA-MDS approach seems prohibitively expensive computationally. One answer is to find a model problem that has an appropriate combination of system size, expected waiting time for occurrence, and the expected duration. It should also be pointed out that there are interesting folding events and transitional events for which simulation is computationally feasible. And, advances in high-performance computing and creative use of existing resources (e.g. [Shi2000]) have extended the range of transitional events accessible to AA-MDS. Furthermore, creative types of IV-AA-MDS simulation such as coupled parallel molecular dynamic (CPMD) simulations and replica exchange molecular dynamics (REMD) simulation have been used to increase the computational resources and the efficiency of IV-AA-MDS in terms of the exploration of phase space. So, AA-MDS continues to be applied in folding and transitional studies.

There are several other courses of action given the computational limitations of AA-MDS that lead to generation of trajectories that are thought to be less accurate, and possibly less detailed, than IV-AA-MDS trajectories. These approaches usually enable application to problems that cannot be studied in adequate detail by AA-MDS. Some of these approaches are described in subsection 3.3.1 and chapter 5. It may be desirable to apply these approximate methods to the study of conformational transitions in smaller systems over shorter time intervals thereby enabling comparisons with conformational transitions generated by AA-MDS. In summary, in spite of computational limitations, AA-MDS has importance in the study of conformational transitions of biomolecules.

2.4.2 Why BV-AA-MDS?

A boundary value approach may not always be an appropriate approach for the study of conformational transitions. It seems appropriate when specific beginning and ending structures are desired, and these structures can be adequately characterized mathematically. But, if these two criteria are not satisfied, an initial value approach may be appropriate.
As previously mentioned, when studying conformational transitions via AA-MDS, it is desirable to obtain an ensemble of transitional trajectories, not just a single trajectory. If this can be accomplished, it may be desirable to classify the trajectories of the ensemble based on some characteristics of the trajectories. In the interest of unbiased sampling of transitional trajectories, an initial value approach with initial velocities that are appropriately scaled and randomly generated from an appropriate statistical distribution (e.g. Gaussian, Maxwell-Boltzmann) has some attractive features. An unbiased sampling approach facilitates empirically derived probabilistic assessments of different events and enables empirical assessments of relative rates based on statistics for different types of transitions. When transitional trajectories are generated by a boundary value approach, probabilistic and statistical assessments are less obvious and relative rates for different types of transitions may not be readily attainable. On the other hand, as will be argued in the next paragraph, a boundary value approach might allow a broader sampling of transitional trajectories for a small sample size.

For beginning and ending conformations from potential energy wells of stable local minima, there may be many different trajectories that go from the beginning to ending conformations. This would be especially likely if different intervals are considered and beginning and ending conformations are designed to approximate an entire well. When trajectories are classified based on important properties of intermediate conformations, the different classifications groups for the trajectories are commonly called pathways. In the boundary value approach, a solution trajectory is likely to be similar to the initial trajectory. If there are multiple pathways between two different conformations, some pathways may be more densely populated than others. So, the probability that an arbitrary transition trajectory follows a certain pathway could vary significantly from by pathway. In this case, with a boundary value approach using a set of initial trajectories that are similar to actual trajectories from relatively rare pathways may result in convergence to trajectories from rare pathways. So, in this way as well, the boundary value approach could provide a broader sampling of trajectories than an initial value approach with a similar sample size, since the initial value approach will tend to provide a majority of solution trajectories from the most probable pathways. Finally, note that if a small sample of trajectories can be generated, methods
described in ‘Transition path sampling’ by Bolhuis et.al.([Bol2002]) could be applied to efficiently produce a larger sample of transition trajectories.

There are differences in the computational costs of IV-AA-MDS and BV-AA-MDS. The computational cost per continuous Newtonian trajectory for IV-AA-MDS is much cheaper than for BV-AA-MDS. But, if transitional trajectories are sought, then it is more reasonable to consider computational cost per continuous transitional Newtonian trajectory. An illustration of differences in computational cost per continuous transitional Newtonian trajectory for an initial value approach and a boundary value approach using the SDE numerical method for is given in [Ole1996] and [Bai2006]. A similar illustration is given in section 2.4.4 of this dissertation for an initial value approach and a boundary value approach using the MS numerical method. Multiple shooting and finite difference methods for BV-AA-MDS both involve an iterative process to minimize an objective function by optimization techniques. In analyzing the effectiveness of initial value and boundary value approaches for finding transitional trajectories, factors of primary importance are the effectiveness and efficiency of these optimization methods and the distribution of the waiting time for a transitional trajectory to occur by chance using an initial value approach. The relevance of the comparison mentioned above is to suggest a means for weighing advantages and disadvantages of different approaches and also show that there may be times when the boundary value approach is preferable. It is hoped that more knowledge about when the different approaches are preferred will be obtained in the future.

Multiple shooting methods could be subclassified so that we consider three classifications of numerical methods for BV-AA-MDS: (1) single shooting, (2) finite differences, and (3) multiple shooting. With this classification, it is noted that multiple shooting combines features of the first two methods. Like the finite difference methods, for multiple shooting a mesh is chosen and an initial guess for values of a solution trajectory at the mesh points is required. In multiple shooting, IVP’s are solved to determine the trajectory between mesh points whereas in finite difference methods, interpolation or collocation methods are used. Discussion about the relationships between the different methods are recurrent in the popular
reference book for numerical solution of BVP’s for ODE’s, *Numerical solution of boundary value problems for ordinary differential equations* by Ascher, Mattheij, and Russell ([Asc1995]) and also in *Numerical analysis* by Stoer, and Bulirsch.([Sto2002]). It is noted parenthetically here that shooting and also multiple shooting are sometimes referred to as ‘initial value’ methods since IVP’s are subproblems in the algorithms designed to solve a BVP. However, the term ‘initial value’ methods will not be used here in an attempt to avoid confusion with initial value AA-MDS.

There are advantages to a boundary value approach for AA-MDS. For the initial value approach to AA-MDS, it has been established that a time step of about $10^{-15}$ s is required in order for a numerical solution to reasonably approximate an analytical and exact solution. Work reported in [Gil1992], [Ole1996] and [Elb1999] provides substantial anecdotal evidence that, when using a finite differences boundary value approach, the time step can be much larger. Using this approach, there is an inverse correlation between the magnitude of the time step and the accuracy of the resulting trajectory, but the resulting trajectory using larger time step can still be a useful approximation. In contrast, using the initial value approach with a similarly large time step results in what amounts to a complete abandonment of the exact trajectory ([Elb1999]). But, how large can the time step be for the boundary approach and how do we assess the accuracy of an approximate trajectory? With respect to the latter questions, in this dissertation, we consider a boundary value solution to be sufficiently accurate if it can be verified to satisfy the equations of motion using an initial value approach with an adequately small time step (a time step of about 0.75 fs was used here), and if it also satisfies the boundary value problem within a specified tolerance level. This requirement essentially means that we seek trajectories that are as accurate as can be expected considering the limitations in accuracy of numerical solution of ODE’s. The former question about the magnitude of the boundary value time step will be revisited in section 2.4.3.

There is another reason for optimism about the boundary value approach. When a transitional trajectory is found, the actual transition time between local minima or between wells can be very fast. However, using random methods of assigning initial velocities, the waiting times for desired
conformational transitions between minima can actually be very long. A significant portion of the simulation time is spent in local minima of the potential energy surface and not in transitions between different type of conformations. As an example, expected waiting times for transitions between wells of the potential energy surface of the alanine dipeptide in solution can be on the order of hundreds of picoseconds or nanoseconds ([Che2004]). Even with a good initial guess, it may be that waiting times for transitions precisely between local minima of a potential energy surface of the alanine dipeptide could actually be much longer than the less specific transitions between wells. In a vacuum, transition times (i.e. actual time from the beginning of the transition to the end) are only 0.1-0.3 picoseconds and in solution, transition times are reported to be similar in [Wou2001]. For larger molecules and more complex conformational changes *in vivo*, it is hypothesized that many conformational transitions may occur as a sequence of rapid and relatively rare transitions between local minima of a potential energy surface. Interjected between these transitional events are potentially long waiting periods spent in these local minima. These waiting periods between transitions can be quite long and can occupy much of the simulation time using an initial value approach. The boundary value approach, it is postulated, will tend to bypass these long waiting periods.

In section 2.2.2, it was argued that equilibrium points in phase space corresponding to local minima of a potential energy surface are Lyapunov stable in theory. Due to energy conserving properties of numerical methods like the velocity Verlet algorithm which will be described in section 3.3.7, an AA-MDS trajectory can exhibit this type of stability for long time periods in practical applications as well. This implies that for initial conditions that are measured perturbations from the equilibrium point of phase space, the system will remain within some measurable distance of that equilibrium point. Yet, an IV-AA-MDS trajectory has been documented to be highly sensitive to small differences in initial conditions (e.g. [Fre2002], [Sch2002], [Rap1995]). [Fre2002] includes an example in which two systems with similar but slightly different initial conditions become ‘nearly uncorrelated’ after about 0.24 ps. This kind of sensitivity might lead to questions about the usefulness of AA-MDS since the numerical solution of
ODE’s always is always subject to some error. Below is a brief description of some relevant research, that adapted from the presentation in [Gil1992].

Suppose there is an operator for computing a Newtonian trajectory exactly that takes a point in phase space, \( y(t_0) \) and integrates it forward by a time \( \Delta t \). Let \( \Phi^{\Delta t} \) represent this mapping on some interval \([t_0, t_f]\) so that

\[
(2.97) \quad \Phi^{\Delta t} y(t_0) \rightarrow y(t_0 + \Delta t).
\]

Suppose there is another operator that only approximates a Newtonian trajectory and is represented by \( \Psi^{\Delta t} \)

\[
(2.98) \quad \Psi^{\Delta t} y(t_0) \rightarrow y_A(t_0 + \Delta t)
\]

Repeated application of \( \Psi^{\Delta t} \) produces an approximation, \( y_A \), to \( y \) on \([t_0, t_f]\). (\( \Phi^{\Delta t} \) is known as the phase flow operator and \( \Psi^{\Delta t} \) is known as a discrete evolution operator. The appendix of this chapter (section 2.6) contains more about these operators and some important results about consistency, stability, and convergence of numerical methods for solving ODE’s.) For all \( t \) such that \( t_0 < t < t_f \), suppose we can assume that

\[
(2.99) \quad \| \Phi^{\Delta t} y_A(t) - y_A(t_0 + \Delta t) \| < \alpha
\]

for some presumably small value, \( \alpha \). Define a sequence with this property to be an \( \alpha \)-pseudotrajectory. Also, define the term ‘\( \beta \)-shadow’ as follows. A trajectory, \( y \), \( \beta \)-shadows a sequence, \( y_A \), on a connected subinterval \( I \subset [t_0, t_f] \) if

\[
(2.100) \quad \| y_A - y \| < \beta
\]

on \( I \). So, a sequence \( \beta \)-shadows another sequence for a period of time if it stays within some distance \( \beta \). The shadowing lemma, then, is as follows:

\[
(2.101) \quad \text{(hypothesis)} \quad \text{For every } \beta > 0, \text{ there is an } \alpha > 0 \text{ such that every } \alpha \text{-pseudotrajectory of a system is } \beta \text{-shadowed for some period of time, } I, \text{ by a true trajectory.}
\]
There is some empirical evidence that the shadowing lemma holds for long time intervals, but a proof of this lemma is lacking ([Fre2002]). Even if the shadowing lemma holds, the practical implications could depend on relationships between values of $\alpha$, $\beta$, and $I$. Still, it could be said that the evidence provides support for the usefulness of IV-AA-MDS.

In section 2.1, it was noted that an equilibrium point of a system governed by (2.1) that corresponds to a local minimum of the potential energy surface is Lyapunov stable. But, the sensitivity to perturbation of initial conditions described earlier in this section is applicable to initial conditions corresponding to an equilibrium point of phase space and this sensitivity suggests some sort of instability.

For the purpose of understanding Lyapunov stability and the sensitivity to initial conditions described in this section, let's focus on an equilibrium point of a system governed by (2.1) that corresponds to a local minimum of the potential energy surface that is Lyapunov stable. Note that this point is Lyapunov stable in theory. But, in AA-MDS, a numerical solution of (2.1) is required. In some cases, it could be that the approximate solution is not Lyapunov stable, especially over long time intervals. In terms of the variables of section 2.1, this could be the case if for the range of values for $\varepsilon$ that are considered relevant, the value of $\delta$ is less that the local error for the numerical algorithm. However, ignoring the error resulting from a numerical solution, it might seem there is a contradiction, but it can be cleared up by considering another type of stability, asymptotic stability. This type of stability is defined as follows:

\begin{equation}
(2.102) \text{(definition)} \quad \text{Let } y^*(t) \text{ be a stable (or uniformly stable) solution of a 1st-order system of differential equations } y'(t) = h(t, y). \text{ If there additionally there exists an } \delta(t_0) > 0 \text{ such that}
\end{equation}

$$
\| y(t_0) - y^*(t_0) \| < \delta \implies \lim_{t \to \infty} \| y(t) - y^*(t) \| = 0$$

then, $y^*(t)$ is asymptotically stable on $t \geq t_0$. ([p. 224 of Jor1983]).

This definition applies for any stable solution. We are focusing on the case where $y^*(t)$ is an equilibrium solution, though, in which case, without loss of generality we may assume that $\| y^*(t) \| = 0$. For a small perturbation from an equilibrium point of an isolated system, the Hamiltonian for the resulting trajectory,
is constant at some positive real amount bounded away from zero. Therefore, the simulation of isolated system subject to the Newtonian equations of motion cannot be asymptotically stable.

Now, how does lack of asymptotic stability relate to the sensitivity to perturbation of initial conditions. Specifying total energy of the system and the initial coordinates for a simulation effectively imposes constraints on the range of possible conformations. For example, conformations with potential energy higher than the initial total energy of the system are not possible unless the numerical method and computer arithmetic lead to a deviation from the initial total energy that is at least as much as the difference between the aforementioned energies. Now, for two simulations with similar initial conditions where one set of initial conditions is an arbitrarily small perturbation of the other at a fixed $t_0$ and $t_f$ goes to $\infty$, we can expect that the distribution of the location in the two systems at time $t_f$ will be asymptotically independent. Depending on the range of accessible phase space they could be independently distributed in a range of similar conformations or they could be independently distributed in a range which includes many different conformations. Regardless, as the length of time of a simulation grows, the point in accessible phase space assumed at the end of the simulation will become more and more sensitive to the initial conditions.

In the next section, section 2.4.3, we will consider stability, magnitude of time step, and computational complexity as we consider in more detail the feasibility of different numerical approaches for large systems and long time intervals. A brief preview is provided here. In general, asymptotic instability can be expected to restrict the usefulness of a single shooting algorithm to shorter time intervals. On the other hand, finite difference methods to produce accurate trajectories involve a prohibitively large number of parameters (atomic coordinates and possibly velocities for $3n$ atoms at each time point). Finite difference methods with long time steps can only be expected to produce approximate trajectories. Since the multiple shooting method involves revision to initial velocities at shooting points based on results from a previous iteration, the Lyapunov instability suggests that there is a limit on the length for subintervals in order to expect effective revisions. But, the multiple shooting algorithm may be a reasonable middle ground between single shooting and finite difference methods. Particularly by using parameter reduction methods, the multiple shooting algorithm offers an approach that may provide a manageable number of
parameters but which is not greatly impacted by Lyapunov instability. Also, the shadowing lemma provides some reason to think that a multiple shooting trajectory that approximately satisfies continuity at shooting points and approximately satisfies boundary conditions may be shadowed by a true trajectory.

2.4.3 Why multiple shooting?

In the previous section, motivation for the BV approach to AA-MDS was provided. Challenging aspects of the three previously mentioned methods for numerical solution of BVP’s for ODE’s— (1) single shooting, (2) finite differences, and (3) multiple shooting — are described in this section.

2.4.3.1 Challenging aspects of single shooting

First, single shooting methods are considered. Since it has been assumed that the AA-MDS force field, \( f(x(t)) \) does not have any stochastic or random terms, the solution to an IVP for AA-MDS theoretically is almost completely determined by the initial conditions. The numerical method and the computer system used to perform the calculation can also affect the solution, but, for a sufficiently short time interval, the effects are expected to be minimal. Now, consider single shooting for the BVP (2.27),(2.28). Also, assume that boundary conditions are defined in terms of absolute position. So, as in the example from section 2.3.1.1, we have the linear boundary conditions, \( x(t_0) = x^0 \) and \( x(t_f) = x^f \) where desired beginning and ending coordinates are given by \( x^0 \) and \( x^f \) and

\[
(2.103) \quad r(y(t_0), y(t_f)) = [x(t_0) - x^0; x(t_f) - x^f].
\]

Recall that \( x(t) \) is a subset of \( y(t) \). The single shooting method has \( 6n \) parameters. If, in the initial iteration, the initial coordinates can be selected to satisfy the boundary conditions at \( t=t_0 \), and can also be fixed in following iterations, then the number of parameters to be determined in the iteration process is effectively reduced to \( 3n \) parameters. In either case, the number of parameters is less than the number of parameters for the other two methods. The single shooting numerical method in this context is designed to find a set of velocities for \( t=t_0 \) that result in a solution to the BVP.
One difficulty that can be encountered in the general application of single shooting is that for some parameter vectors, a solution to the resulting single shooting IVP might not even exist on the entire time interval for the BVP. Still, a solution to the BVP could exist and be attained by IV-AA-MDS if the appropriate parameter vector is chosen. If a parameter vector for which a solution to the IVP does not exist on the entire time interval of the simulation is chosen as an initial parameter vector or is the parameter vector for a later iteration, then the single shooting method will fail. This difficulty of the single shooting method is discussed in [Sto2002] and [Asc1995]. Multiple shooting methods and finite difference methods can be used to alleviate this difficulty. In [Bai2006], simple biomolecular examples relevant to BV-AA-MDS are studied. The authors produce evidence to suggest that for the examples they study, there are many trajectories, many of which are very similar, that satisfy specified boundary conditions on a wide range of time intervals. Of course, this does not prove that solutions always exist. But, it suggests that, it is likely that if appropriate time intervals, appropriate initial guesses, and appropriate numerical methods are used, that it is likely that solutions can be found.

It was reported in [Fre2002] indicated that tiny changes in initial conditions can lead to trajectories that are almost uncorrelated after about 0.25 \( \text{ps} \) of simulation time. How does this relate to the effectiveness of the single shooting method? Consider the effects of sensitivity to initial conditions on the single shooting method implemented with a globally convergent modification of Newton’s method. It would seem that a minimal requirement for the use of the single shooting method is that the step adjustment, \( \xi_k \), to the parameter vector, \( s_k \), for the next iteration can be computed reliably. Based on the global optimization methods described above, a reliable calculation of \( \xi_k \) would require a reliable calculation of \( \xi^{Nwtn(k)} \) and possibly \( \xi^{Cchy(k)} \). Upon inspection of the formulas for \( \xi^{Nwtn(k)} \) and \( \xi^{Cchy(k)} \) given in (2.76) and (2.78), respectively, it seems reasonable to think that the condition number of \( F'(s) \) might be a good indicator of the reliability of the calculations for \( \xi^{Nwtn(k)} \) and \( \xi^{Cchy(k)} \). The condition number, \( \kappa \), of \( F'(s) \) can be computed at each time point of an initial value simulation. The condition number of a matrix, \( A \), is computed using the formula \( \kappa(A) = ||A|| ||A^{-1}|| \) where the notation \( ||A|| \) represents the norm of the matrix \( A \). There are different methods for determining the norm...
of a matrix and different corresponding formulas for the norm of a matrix. The value of the condition number may vary depending on the choice of norm. In this work, we use the $L_2$ norm. The definition for this matrix norm is somewhat intuitive. For an $n \times n$ matrix $A$, it is the maximum value of $\|Au\|_2 / \|u\|_2$ for all vectors $u$ of size $n \times 1$. For solution of a BVP using single shooting (or multiple shooting), the condition number of $F'(s)$ is of particular interest. The reader is referred to Chapter 2 of [Dem1999] for an introduction to the topic of matrix norms.

A condition number can be computed for all $t$ between $t_0$ and $t_f$. In general, this pointwise condition number of $F'$ depends on the parameter vector, $s$, of the initial value simulation, and $F'$ varies with time. With these dependencies in mind, here we write $\kappa(F(t;s))$. To provide an idea of how $\kappa(F(t;s))$ varies with time and energy level, thirty different initial value simulations for alanine dipeptide in vacuo have been performed assuming absolute boundary conditions of the form of (2.6). There are ten simulations for each of three different sets of initial coordinates and initial velocity directions. The initial velocity directions are chosen as directions which have been previously determined to, at some scale, or energy level, produce a trajectory that transitions from one type of conformation to another. With respect to Figure 1.2, Figure 1.3 and the corresponding description of alanine dipeptide conformations in section 1.3, the three conformational transitions can be categorized as $C7_{eq} \rightarrow C6$, $C7_{eq} \rightarrow C5_{\beta}$, and $C7_{ax} \rightarrow C7_{eq}$. For each of the three initial coordinate/initial velocity direction pairs, ten different initial velocity scaling factors are used. The different scaling factors correspond to different (total) energy levels, so these experiments provide a way to inspect the possible relationships between energy level and condition number. The time step for these simulations is $\Delta t = 0.015725 \text{ AKMA units} \approx 0.769 \text{ fs}$. In each simulation, 9999 steps are performed to give a simulation time of $t_{\text{max}} = (9999)(0.015725) \text{ AKMA units} \approx 157.24 \text{ AKMA units} \approx 7.69 \text{ ps}$. In Figure 2.3, Figure 2.4, and Figure 2.5, $\kappa(F(t;s))$ is shown as a time series (in green) in subplots for each IVP simulation using a logarithmic scale for simulations for initial direction for transition from $C7_{eq}$ to $C6$, $C7_{eq}$ to $C5_{\beta}$, and $C7_{eq}$ to $C7_{ax}$, respectively. The time series (blue) in these subplots give corresponding values of $\kappa(Y_0(t;s))$. As described in [Deu2002], $\kappa(Y_0(t;s))$ is defined to be the pointwise condition number of the initial value problem. In Figure 2.3, the graphs are shown, reading from left to
right, then, top to bottom, for simulations with total energy in \( \text{kcal mol}^{-1} \) of \(-11, -15, -10, -5, 0, 5, 10, 15, 20\) and \(25\). In Figure 2.4 and Figure 2.5, similar graphs are shown. Energy values in \( \text{kcal mol}^{-1} \) are \(1, -15, -10, -5, 0, 5, 10, 15, 20\), and \(25\) for the subplots of Figure 2.4. and energy values in \( \text{kcal mol}^{-1} \) are \(73, -10, -5, 0, 5, 10, 15, 20, 30,\) and \(40\). In each of the three figures, the 1\text{st} subplot corresponds to an energy level that was originally found to result in the indicated transition. The remaining subplots are a series of incremental rescalings of the initial velocity of the original trajectory. (These time series also have relevance with respect to multiple shooting as will be discussed in section 2.4.3.6.)

Each of the three figures (Figure 2.3, Figure 2.4, and Figure 2.5) indicates that \( \kappa(F'(t;s)) \) tends to increase exponentially (the y-axis is scaled logarithmically) as \( t \) increases. Interestingly, there are a few low energy simulations (\(-15 \text{ kcal}\) and \(-10 \text{ kcal}\)) in Figure 2.4 in which the increase is only small on the logarithmic scale at least through the ending time of 7.7 ps of the simulations. A possible explanation for these smaller increases is that \( \kappa(F'(t;s)) \) will, in general, tend to grow more slowly at a function of time at lower total energy levels where the accessible phase space is limited (see argument in section 2.4.2). The overall trend is still clear. Three questions to ask are the following:

1. What is the level of condition number for which the solution to the single shooting linear system becomes unreliable?

2. Is the nature or rate of the tendency to increase correlated with the total energy level in the system?

3. What is the number of steps that we can expect to be able to take in an initial value simulation before the Jacobian, \( F'(s) \), becomes ill-conditioned?

While each of these questions probably warrants further study for a definitive answer, we can at least attempt to ascertain what the data in Figure 2.3, Figure 2.4, and Figure 2.5 appear to suggest.

Regarding question 1, we make the assumption that if the relative error in the solution vector \( \xi \) is below \(10^{-4}\), then \( \xi \) will be reliable. We then require that the relative error in estimating \( \xi \) is below \(0.5 \times 10^{-4}\).
To insure this, we require that the product of the condition number and the machine epsilon be below $0.5 \times 10^{-4}$. For 32-bit arithmetic using double precision in MATLAB, machine epsilon is $2.2 \times 10^{-16}$. Using this computing environment, the condition number must be less than $\text{TOL}_\kappa = 0.5 \times 10^{-4} / 2.2 \times 10^{-16} = 2.27 \times 10^{11}$ for $\xi$ to be reliable.

Regarding questions 2 and 3, we refer to Figure 2.6. In this figure the source data of Figure 2.3, Figure 2.4, and Figure 2.5, are used. For each of the 30 IVP’s, the 1st step for which $\kappa(F'(t; s))$ exceeds $TOL_\kappa$ is recorded. In Figure 2.6, this step number is plotted versus the total energy level of the simulation. Regarding question 2, there is some indication that $\kappa(F'(t; s))$ tends to increase at a faster rate as a function of time for higher energy levels, although, in the energy scale of these simulations, correlations between rate and energy level are not so strong. Regarding question 3, in none of the simulations does the condition number exceed $\text{TOL}_\kappa$ in the first 2000 steps $\approx 31.44 \text{ AKMA units} \approx 1.54 \text{ ps}$. An objective for future studies will be to develop a better understanding of the factors influencing maximum time interval for single shooting.

The observations that have been made above suggest there is an important limitation for length of the total time of simulation when using the single shooting method. Another limiting aspect of the single shooting method as described above is the computational cost of the $6n \times 6n$ linear system solution of (2.45), which is on the order $O(n^3)$. So, the computational cost grows as a cubic factor of the number of atoms. Using standard methods for the linear system solution, this cost becomes problematic for $n > 350$. However, using approximate approaches to this linear system solution, it is possible to apply the single shooting method (or multiple shooting method) with much larger values for $n$. One approximation method is described in Chapter 4. Additionally, the computational cost of shooting methods will be studied in further detail in section 2.4.4.

As mentioned above, for a BVP, $\kappa(F(t_0; s))$ and $\kappa(F(t_f; s))$ are of particular interest. Assuming the numerical solution of the BVP is computed strictly in the forward direction, then $\kappa(F(t_0; s)) = 1$. The pointwise condition number for the IVP,
\[ y'(t) = h(y(t)), \quad t_0 < t < t_f \iff x''(t) = f(x(t)), \quad t_0 < t < t_f \]

(2.105) \quad y(t_0) = s \iff x(t_0) = x_0, \quad v(t_0) = v_0

is \( \kappa(Y(t;s)) \) where

\[ Y(t) = \frac{\partial y(t;s)}{\partial s} \]

is the solution of

\[ \frac{d}{dt} Y(t;s) = \frac{\partial}{\partial y} h(y(t,s)) Y(t;s), \quad t_0 < t < t_f \]

The IVP condition number for the interval \([t_0, t_f]\) is defined as

\[ \kappa(Y(s), [t_0, t_f]) = \max_{t_0 < t < t_f} \kappa(Y(t;s)) \]

As described in [Deu2002], we can define the single shooting IVP condition number for the interval \([t_0, t_f]\) as

\[ \kappa(F'(s), [t_0, t_f]) = \max_{t_0 < t < t_f} \kappa(F'(t;s)). \]

While it is not true in general that \( \kappa(F'(s), [t_0, t_f]) = \kappa(F'(t_f;s)) \), these plots suggest that \( \kappa(F'(t;s)) \) tends to be approximately the same order of magnitude as \( \kappa(F'(t_f;s)) \). Finally, there are some differences between the graphs of \( \kappa(F'(t;s)) \) and \( \kappa(Y_d(t;s)) \), but, overall, they are quite similar. This may not be surprising since \( F'(t;s) \) is largely determined by \( Y_d(t;s) \). This correlation suggests a link between the sensitivity to perturbations of initial conditions for an initial value problem and the sensitivity to perturbations of initial conditions for the single shooting method applied to a boundary value problem. In summary, the exponentially increasing growth of \( \kappa(F'(t;s)) \) for the single shooting method for BV-AA-MDS is due to the lack of asymptotic stability and the fact all points in accessible phase space for an isolated system are equally likely at time \( t \) as \( t \to \infty \).

2.4.3.2 Challenging aspects of finite difference methods

In the previous subsubsection, we concluded that for longer time intervals, the Jacobian, \( F'(s) \), for the single shooting method is expected to be ill-conditioned which, in turn, is expected to render the single
shooting method inefficient, and, for practical purposes, ineffective. This ill-conditioning doesn’t necessarily imply that the BVP itself is ill-conditioned. We noted above that $\kappa(Y(t,s))$ and $\kappa(Y(s),[t_0,t_f])$ are the pointwise condition numbers and interval condition numbers for the initial value problem (2.104),(2.105). However, it is NOT true that $\kappa(F'(t,s))$ and $\kappa(F'(s),[t_0,t_f])$ are the pointwise condition numbers and interval condition numbers for the BVP (2.27),(2.28). These numbers reflect the conditioning of the single shooting method, but, not necessarily the BVP itself. Formulas for a pointwise condition number, $\rho(t)$, and an interval condition number, $\rho([t_0,t_f])$ are their derivations are given in Theorem 8.5 of [Deu2002], pages 395 and 396.

Application of the theory of conditioning of nonlinear BVP’s is a intricate and rather complicated subject that will not be discussed in detail here. It is relevant, however, to highlight some observations from Chapter 8 of [Deu2002] about the condition number of a BVP and the condition number of finite difference methods. To begin, below is an example that illustrates the potential for sharp contrasts between the magnitude of $\rho(t)$ and $\kappa(Y(t,s))$.

### 2.4.3.3 Example 6: Conditioning of scalar linear 2nd-order BVP and IVP

Example 8.4 on page 395 of [Deu2002] gives an example of a simple linear scalar 2nd-order BVP for which $\rho(t)$ is constant but for which $\kappa(Y(t,s);t,s)$ for a corresponding IVP grows exponentially. The BVP

\[
(2.110) \quad x'' - \lambda^2 x = 0, \quad x(t_0) = x_0, \quad x(t_f) = x_f, \quad t_0 < t < t_f
\]

and corresponding IVP

\[
(2.111) \quad x'' - \lambda^2 x = 0, \quad x(t_0) = x_0, \quad x'(t_0) = v_0, \quad t_0 < t < t_f
\]

have pointwise condition numbers $\rho(t) = \lambda$ and $\kappa(Y(t,[x_0,v_0])) = \lambda \exp[\lambda(t_f-t_0)]$END OF EXAMPLE

It is useful to have knowledge about $\rho(t)$ for AA-MDS, that is the conditioning properties of BV-AA-MDS. For the many dimensioned AA-MDS with its nonlinear, multi-scaled force field, analytical
calculation of $\rho(t)$ is elusive. One could use the shadowing lemma as a starting point for an argument that BV-AA-MDS is, at least, in some cases, well-conditioned as the following example illustrates.

### 2.4.3.4 Example 7: Shadowing lemma and conditioning of BVP

Consider a theoretically exact algorithm introduced in section 2.4.2 of the form

\[(2.112) \quad u_1, u_2, ... \text{ where } u_{i+1} \rightarrow \Phi u_i \text{ for } i = 1, 2, ...\]

Also, consider the algorithm

\[(2.113) \quad \bar{u}_1, \bar{u}_2, ... \text{ where } \bar{u}_{i+1} \rightarrow \Phi \bar{u}_i \text{ for } i = 1, 2, ...\]

where

\[(2.114) \quad \|\bar{u}_i - u_i\| < \alpha\]

Then, the sequence (2.113) is clearly an $\alpha$-pseudotrajectory. If the shadowing lemma is applicable, then we have that for some $\beta > 0$, is $\tilde{u}$ is $\beta$-shadowed for some period of time by the exact trajectory $u$. But, $\tilde{u}$ is also a true trajectory differing from $\tilde{u}$ in that initial coordinates are perturbed. For $\beta$ sufficiently small, we have that small changes in boundary conditions defined by initial and ending coordinates lead to small changes in a solution trajectory over some period of time. Furthermore, for $\beta$ at some pre-determined threshold for a well-conditioned problem, the maximum value length of time for which $\tilde{u}$ is $\beta$-shadowed by \{u_i\} indicates the length of time for which a BVP remains well-conditioned. END OF EXAMPLE

The shadowing lemma itself is only based on empirical evidence, but even if it is always true, the actual values of $\alpha$, $\beta$, and $t_f$ could be chosen to fit a wide range of BVP’s from the very ill-conditioned to the very well-conditioned. If a BVP is well-conditioned, it would be desirable to use a numerical method that is conditioned similarly. To this end, there exist some relevant theoretical results about finite difference methods. In finite difference methods for linear BVP’s, there is a system of linear equations to solve to arrive a solution. In [Asc1995], it is asserted that for a linear BVP on interval $t_0 < t < t_f$ with condition number $\rho(t_f - t_0)$, the linear system solve can be accomplished in a way that ensures that the
condition number of the matrix of that linear system solve is less than \((D+1)\rho(t_f-t_0)\) where \(D+1\) is the number of elements in the mesh for the finite difference method. For nonlinear BVP’s some theoretical results can be obtained which allow a similar conclusion under certain conditions. Unfortunately, the conditions would be difficult to establish or verify in the case of BV-AA-MDS. Still, with respect to conditioning, it seems reasonable that finite difference methods would tend to be advantageous, in general, in comparison with single shooting methods, which are more sensitive to Lyapunov instability. This observation would be consistent with arguments developed from different perspectives in [Gil1992], [Ole1996], and [Elb1999].

An important consideration for a finite difference method is mesh selection. In [Ole1996] and [Elb1999], a finite difference method, the SDE method, was introduced as a method to find trajectories that satisfy boundary conditions of AA-MDS BVP’s but may only approximately satisfy Newton’s equations of motion. A uniform mesh selection is used with different time steps. The density of the mesh is directly correlated with the accuracy of the approximation. A goal of this dissertation is to explore and develop methods to find trajectories that not only satisfy boundary conditions of AA-MDS BVP’s but also satisfy Newtonian equations of motion (within a level of tolerance to allow for error in numerical solution of differential equations and numerical arithmetic). With these things in mind, it is logical to inquire about the maximum value for the time step, \(\Delta t\), for the finite difference methods described in section 2.3.3.

The most important degrees of freedom for the determination of the conformation of a protein consist of the sequential pairs of dihedral angles around the C\(^\alpha\) atoms of the individual amino acids that make up the protein. For small proteins, it is asserted in [Wou2001] that substantial changes in dihedral angles corresponding to conformational changes from one locally stable state to another can occur over a time period of less than 0.05 ps. Protein conformation changes that occur over long time intervals can contain short subintervals with rapid changes with long interludes in metastable states with minor local conformational fluctuations but no significant overall conformational changes. These metastable states could be described mathematically as wells of the potential energy surface corresponding to local minima.
It is logical to assume that, for an accurate depiction of the process of changing conformations, it is necessary to include snapshots of a solution trajectory with a frequency below than 0.05 ps, no greater than, for the sake of argument, say 0.05 ps /10 = 0.005 ps. For Δt = 0.005 ps is a larger time step by a factor of about 2.5 to 5 when compared to unconstrained or bond length constrained initial value dynamics.

The estimated maximum value for Δt = 0.005 ps for finite differences is larger than Δt for initial value problems. In the finite difference BVP approach, a system of nonlinear equations is produced for each iterations and the total number of variables will be equal to \( 6nN = 6n(t_f-t_0)/\Delta t. \)

### 2.4.3.5 Example 8: Number of variables for application of finite differences

With this observation, we see that for moderate values, say \( n = 253 \) and \( (t_f-t_0) = 1 \, \text{ns} \), we would have \( 6n(1 \, \text{ns}) / 0.005 \, \text{ps} = 304,800,000 \) variables! While there may be ways to reduced the number of variables in practice, and while approximation solutions to the linear equation that result may be possible, the potentially large number of variables associated with usage of finite difference methods could be a significant drawback. END OF EXAMPLE

### 2.4.3.6 Challenging aspects of multiple shooting methods

In section 2.4.3.1, we observed that the single shooting method was well-conditioned for \( t_f - t_0 < 1.5 \, \text{ps} \) and we also observed the equivalent orders of magnitude of \( \kappa(Y(t,s)) \) and \( \kappa(F'(t;s)) \). These results can be used to infer that the IVP’s on the subintervals for the multiple shooting method would be well-conditioned given that \( (t_{i+1} - t_i) < 1.5 \, \text{ps} \), for \( i = 0,1,2,\ldots,n-1 \). One might assume that the multiple shooting method are well-conditioned as long as the IVP’s on the subintervals are well-conditioned. (This may not be always be true, but probably is often true). In comparison with the finite differences method, then, the multiple shooting method then may be implemented with a smaller number of variables. And, in comparison with single shooting, it can be implemented over much longer time intervals while still retaining a manageable condition number.
2.4.3.7 Example 9: Conditioning, variable count for application of multiple shooting

As an example, we consider the example of the previous section, \( n = 253 \), \( (t_f - t_0) = 1 \text{ ns} \). If multiple shooting methods were applied with 667 subintervals so that the length of a subinterval was 1.5 ps, then there are 1,012,000 variables. This is still a large number of variables but it is \( 1.5/0.005 = 300 \) times less than the number of variables using the finite differences method. Single shooting over 1 ns is impractical as the method would surely become ill-conditioned. END OF EXAMPLE

Straightforward application of multiple shooting methods for an example of this type would certainly represent a challenging optimization problem. Methods for simplifying boundary conditions, parameter reduction, and possibly also reducing the total time interval of simulation will be discussed in later chapters. But, simply stated, the multiple shooting method represents a compromise between single shooting and finite differences that may keep the problematic aspects of those two methods manageable.

2.4.4 Relative computational cost of BV-AA-MDS and IV-AA-MDS

Below, we provide a sketch of how one might compare boundary value and initial value approaches to simulation of transitions between known conformations using distributed computing in terms of computational complexity. Methods for generating an initial trajectory for the multiple shooting method may be different than methods for generating an initial trajectory for an initial value approach, but for simplicity, let’s assume that the value and cost of generating an initial trajectory for these two approaches are approximately equal. So, we assume we have a set of initial trajectories for the multiple shooting method and for an initial value method. Now, define the following variables:

\[ n \quad : \text{number of particles in the system (as previously defined)} \]
\[ T \quad : \text{number of steps for single IVP trajectory} \]
\[ O(\Lambda) \quad : \text{order of operations for force field evaluation} \]
\[ \Pi \quad : \text{number of processors} \]
\[ Z \quad : \text{reduction factor on multi-processor speedup for BVP approach} \]
\[ \mu \quad : \text{expected waiting time for transition (in units of IVP trajectories)} \]
$N$: number of multiple shooting subintervals (as previously defined)

The multiple shooting BVP approach requires calculation of second derivatives of the potential energy function and the Jacobian or approximate Jacobian of the MS linear system, as well as a solution to the MS $6nN \times 6nN$ linear system. We define the additional variables:

- $\Theta$: factor to multiply by $O(\Lambda)$ to account for calculation of second derivatives and calculation of the Jacobian of the MS systems of equations.
- $\Omega$: factor to multiply by $O(\Lambda)$ to account for approximately solving the MS linear system (i.e. the approximation to the linear system solve must be $O(\Omega\Lambda)$)
- $\Xi$: average number of minimization steps in BVP approach.

If $F'$ is dense, solving this system costs $O((Nn)^3)$, but $F'$ is almost block diagonal and efficient methods for this structure cost $O(Nn^3)$. The cost to generate an IVP trajectory is $O(\Lambda)$ where typically, $O(n) < O(\Lambda) \leq O(n^2)$. The numerator of the formula for $\rho$ given below is an estimate of the computational complexity for finding a conformational transition by IVP methods and the denominator is an estimate of the complexity for a MS BVP method. So, to compare the two approaches from this perspective, we see that larger values of $\rho$, specifically values greater than 1, suggest preference for the BVP approach.

$$\rho = \frac{\frac{\pi}{\pi}}{\frac{\pi}{\pi}} = \frac{Z\mu}{Z\mu}$$

Can we approximately solve the MS almost block diagonal linear system in $O(\Omega\Lambda)$ and get meaningful solutions? To be able to answer yes to this questions would seem to be a critical challenge for application of this approach for large systems. The results presented here suggest that the number of boundary conditions and the number of parameters necessary to solve AA-MDS BVP’s must be greatly reduced from the full set of $6n$ boundary conditions and full set of $6nN$ parameters. Reductions like these would seem to be an important component of finding approximate solutions that can be accomplished in
\(O(\Omega \Lambda)\). Regardless, we note that a capped 16-residue 253-atom C-terminal \(\beta\)-hairpin from Protein G is of manageable size, so that the challenges described above are not be problematic for systems of this size even without reduction of boundary conditions and number of parameters.

### 2.4.5 Comparison with Stochastic Difference Equation approach

Ron Elber’s Stochastic Difference Equation (SDE) approach, summarized in section 2.3.7, is well-known as an BV-AA-MDS. In this subsection, we focus on relations between the SDE method using time as the independent variable (SDET method) and the MS method. For the purpose of relating SDET and MS, let’s consider the application of the SDET method with a small step size, say \(\Delta t = 0.01\ AKMA\) units over a small time span of 0.03 \(AKMA\) units. So, we have 3 time steps. Furthermore, assume we have a one-dimensional system where one particle, \(x\), of unit mass is subjected to a force field, \(f(x)\), with boundary conditions, \(x(0) = x_0^*\) and \(x(0.03) = x_3^*\).

Let’s consider SDET for this BVP using the Verlet algorithm with an initial velocity of \(v(0) = v_0\).

For notational convenience, let \(x_\tau = x(10^{-2}\tau)\) for \(\tau = 1, 2, 3\). Then, we have

\[
x_1 = x_0 + \Delta t v_0 + \frac{1}{2} \Delta t^2 f(x_0)
\]

And, for \(\tau = 1\) and \(\tau = 2\), we have

\[
x_i = \Delta t^2 f(x_{i-1}) + 2x_{i-1} - x_{i-2}.
\]

This leads to the following nonlinear system of equations:

\[
F^{SDET}(x) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 \\ 0 & 1 & -2 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} -x_0^* \\ -\Delta t^2 f(x_1) \\ -\Delta t^2 f(x_2) \\ -x_3^* \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}
\]

For the SDET method, the optimization problem is

\[
\min_x \left\| F^{SDET}(x) \right\|_2^2
\]
Including formulas for velocity in the system of nonlinear equations, we can write:

\[
\tilde{F}(x) = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 \\
1 & -2 & 1 & 0 & 0 & 0 \\
0 & 1 & -2 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & -1 & 0 & 2\Delta t & 0 \\
0 & 1 & 0 & -1 & 0 & 2\Delta t
\end{bmatrix} \begin{bmatrix}
x_0 \\
x_1 \\
x_2 \\
x_3 \\
v_1 \\
v_2
\end{bmatrix} + \begin{bmatrix}
x_0^* \\
-\Delta t^2 f(x_1) \\
-\Delta t^2 f(x_2) \\
-x_3^* \\
0 \\
0
\end{bmatrix} = 0
\]

Since \( v_1 \) and \( v_2 \) have explicit formulas in terms of \( x_0, x_1, x_2, \) and \( x_3 \), a sequence of conformations corresponding to a solution of (2.119) is a sequence of conformation that also corresponds to a solution of (2.121)

\[
\min_x \left\| \tilde{F}(x) \right\|_2^2.
\]

Now, if we consider converting from a second order system and then applying the velocity Verlet algorithm, we end up with the following system of equations

\[
\tilde{F}(x) = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 \\
1 & -1 & 0 & \Delta t & 0 & 0 \\
0 & 1 & -1 & 0 & \Delta t & 0 \\
0 & 0 & 1 & 0 & 0 & \Delta t \\
0 & 0 & 0 & 1 & -1 & 0 \\
0 & 0 & 0 & 0 & 1 & -1
\end{bmatrix} \begin{bmatrix}
x_0 \\
x_1 \\
x_2 \\
v_0 \\
v_1 \\
v_2
\end{bmatrix} + \begin{bmatrix}
x_0^* \\
\Delta t^2 f(x_0) / 2 \\
\Delta t^2 f(x_1) / 2 \\
\Delta t^2 f(x_2) / 2 - x_3^* \\
\Delta t(f(x_0) + f(x_1)) / 2 \\
\Delta t(f(x_1) + f(x_2)) / 2
\end{bmatrix} = 0
\]

Consider applying MS using the velocity Verlet algorithm for the same BVP with \( \Delta t=0.01 \) AKMA units and \( N = 3 \). The number of subintervals is equal to the number of steps. The system of equations to solve takes the form
In the final form, we see that the system of equations for MS is identical to the system of equations of (2.122) since the only difference between (2.122) and (2.123) is the ordering of the variables and the ordering of the equations. For the dogleg trust region optimization method, the objective function is

\[
\min_x \left\| F(x) \right\|_2^2
\]

Let’s attempt to summarize in a larger context. In section 2.3.3, a finite difference method is described for the BVP (2.1),(2.3). Upon describing the SDET method in section 2.3.7, it was argued that the objective function for SDET of [Ole1996] is equivalent to a finite difference method implemented with a global convergence scheme using an commonly used objective function. In section 2.4.2, reasons to consider BV-AA-MDS are supplied. And, in section 2.4.3, reasons to consider MS as a numerical method are articulated. In this section we have attempted to illustrate the important aspects of the relation between the MS method and the SDET method using an example.

The illustration can be best understood by focusing on three elements of these methods. First, The two different tiers of the mesh are important. The MS method has the mesh, \( \Delta s \), of shooting points and another mesh, \( \Delta t \), that is the union of the mesh used for solving the IVP’s on the subintervals. For finite
difference methods like SDET, there is only one mesh. In standard finite difference applications, we would expect that that the density of a finite difference method would be somewhere in between the density of the two MS meshes, but probably tend to be closer to the density of $\Delta t$. SDET is used, however, as a large-time step method, so let’s assume that the mesh for SDET is comparable $\Delta s$. For the case of uniform distributed subintervals, this implies that $\Delta t = N/(t_f - t_0)$. (Note that $N = \Delta t(t_f - t_0)$). The remaining two elements indicate differences between current implementations. First, there is the method for solving the IVP’s to connect the subintervals. In the MS method we have described, the velocity Verlet algorithm is used. In the SDET method, the central Verlet algorithm is used. Finally, different global convergence schemes are used. In the description of MS methods applied here, dogleg trust-region and damped Newton methods are used. The SDET algorithm has been applied with different methods such as multigrid, conjugate gradient, simulated annealing, and Kaczmarz iterations. It is worth noting that, for either finite difference methods or MS methods, these last two elements could be modified. For both finite difference and MS methods, it may be interesting to consider different methods for solving IVP’s and different global convergence schemes.

In this example, $N = \Delta t(t_f - t_0)$. For $N \ll \Delta t(t_f - t_0)$, how can we conceptually contrast MS with $N+1$ uniformly spaced shooting points and SDET with a uniform mesh with $N+1$ points (including the endpoints)? The MS method involves an attempt to find numerical solutions to a system of ODE’s on each subinterval between shooting points, so that the ODE is solved over the entire interval. The SDET method involves an interpolation between mesh points and an attempt to solve a discretized system of ODE’s that results from considering only the mesh points.

2.5 Summary

The topic of this dissertation is the use of a numerical method—multiple shooting—to solve mathematical problems—BVP’s for ODE’s—which serve as models for important biomolecular phenomena—conformational transitions. This chapter provided an introduction to the multiple shooting method, to some important aspects of the phenomena of conformational transitions of biomolecules, and to
relevant mathematical characterizations of these phenomena. Additionally, finite difference methods and
the stochastic difference equation method have been described to facilitate comparative analysis and
discussion. Some existing work — concepts, terminology, definitions, and hypotheses — has been
summarized here to provide a compact presentation of the foundations for the mathematical models that are
used. The existing work has been supplemented by some original work. In particular, an attempt has been
made to define the well of a potential energy surface as it related to molecular dynamics simulation.

Motivations for the use of AA-MDS, BV-AA-MDS, and the application of MS to BV-AA-MDS
were provided. Because quantum mechanical simulation is computationally prohibitive for large
macromolecules, because classical mechanics is thought to be a good approximation for large
macromolecules, and because experimental approaches don't provide the desired detail that can be obtained
from simulation, AA-MDS has relevance as a primary means to study the dynamics of macromolecules.
The boundary value approach to AA-MDS is reasonable when trajectories with two known conformations
are desired. If appropriate numerical methods are used, the boundary value approach to AA-MDS is not as
severely affected by sensitivity to initial conditions as the initial value approach to AA-MDS. The effect of
sensitivity to initial conditions on the application of the multiple shooting method can be measured by
computing the condition number of the Jacobian for system of nonlinear equations of the MS numerical
method.

The usefulness of the single shooting method is limited by asymptotic instability. The finite
difference method applied to a problem with a large number of particles in the system or a long duration of
simulation time results in an excessively large number of parameters. The MS method can be seen as a
compromise between the finite difference method and the single shooting method, that may keep the
problematic aspects of these other methods manageable. An approach to compare the computational cost
of IV-AA-MDS and BV-AA-MDS using MS for the study of conformational transitions was provided.
Additionally, an illustrative example was included that elucidated similarities and differences between the
SDET method and the MS method. And, the concept of distance matrix interpolation, an approximate method for studying conformational transitions, was provided by supplying a couple of simple examples.

### 2.6 Appendix: Order of accuracy of selected numerical methods for solution of 2nd-order Hamiltonian systems

In this appendix, some formalism is supplied for the operator notation of section 2.4.2. An important result that relates consistency error, or local error, and discretization error, or global error, is stated. And, the local order of accuracy is verified for the methods for numerical solution of 2nd-order Hamiltonian systems that are used or discussed in this work. (Hamiltonian systems are physical systems governed by a 2nd-order differential equation in which the forces are not dependent on the velocities.) These methods are the central Verlet algorithm, the non-central Verlet algorithm, and the velocity Verlet algorithm.

#### 2.6.1 Phase flow, discrete evolution, consistency, and convergence

The function on the right side of (2.27), is a mapping \( h: \Omega_0 \rightarrow \mathbb{R}^d \) where \( \Omega_0 \subset \mathbb{R}^d \) and \( d = 2(3n) = 6n \). As was described more informally in section 2.1, \( \Omega_0 \) is called the phase space, or state space. The independent variable \( t \) must satisfy \( t \in \mathbb{R} \), so the augmented phase space is defined as the set \( \Omega = \mathbb{R} \times \Omega_0 \).

Suppose that \( y(t) \) is the solution to the IVP (2.27), \( y(t_0) = y_0 \). The phase flow, \( \Phi' \), is defined by

\[
(2.125) \quad y(t) = \Phi'^{t-t_0} y(t_0)
\]

for all \( t \) such that \( t_0 < t < t_f \). As described in section 2.3.3, the mesh, represented by \( \Delta' \), is a set consisting of \( D+1 \) points, \( t_0, t_1, \ldots, t_{D-1}, t_D \) that satisfy \( t_0 < t_1 < \ldots < t_{D-1} < t_D = t_f \).

One way to classify numerical methods for solving IVP’s for ODE’s is by the number of previous steps required to advance the solution. A one-step method requires only the value of \( y_{\Delta}(t) \) to determine \( y_{\Delta}(t+\Delta t) \). A multistep method requires the current value, \( y_{\Delta}(t) \), and previous values. For example a two-step method (with a constant time step, \( \Delta t \)) requires \( y_{\Delta}(t-\Delta t) \) and \( y_{\Delta}(t) \) to determine \( y_{\Delta}(t+\Delta t) \).
To approximate $y(t)$, we consider algorithms that construct a mesh function $y_\Delta : \Delta \rightarrow \mathbb{R}^d$. The mesh function is to be generated recursively, i.e.

$$y_\Delta(t_0) = y_0 \mapsto y_\Delta(t_1) \mapsto y_\Delta(t_2) \ldots \mapsto y_\Delta(t_{\Delta t}) \mapsto y_\Delta(t_{\Delta t}) = y_\Delta(t_f)$$

Now, suppose that there is a mapping $\Psi_\Delta t$ such that $\Psi_\Delta t y(t)$ is defined for all $(t,y) \in \Omega$ and for $\Delta t$ sufficiently small. Then, $\Psi_\Delta t$ is known as a discrete evolution. One-step methods can be summarized as follows:

$$y_\Delta(t_0) = y_0$$
$$y_\Delta(t + \Delta t) = \Psi_\Delta t y_\Delta(t)$$

The consistency error, or local error, is defined, for sufficiently small $\Delta t$, as

$$e(t, y; \Delta t) = \Phi_\Delta t y(t) - \Psi_\Delta t y(t)$$

The mesh error is defined as

$$e_\Delta : \Delta \rightarrow \mathbb{R}^d, e_\Delta(t) = y(t) - y_\Delta(t)$$

and

$$||e_\Delta||_\infty$$

is called the discretization error. Assuming that a constant time step, $\Delta t$, is used, a mesh function $y_\Delta$ converges to a mapping $y$ if $||e_\Delta||_\infty \rightarrow 0$ as $\Delta t \rightarrow 0$. The convergence is of order $p$ if $||e_\Delta||_\infty = O(\Delta t^p)$ as $\Delta t \rightarrow 0$.

The consistency is of order $p$ if $e(t, y; \Delta t) = O(\Delta t^{p+1})$ as $\Delta t \rightarrow 0$ locally uniformly in $\Omega$. An important result given as Theorem 4.10 of [Deu2002] states that under certain conditions, the convergence order is the same as the consistency order.

### 2.6.2 Central Verlet algorithm: position

The position for the central Verlet algorithm is given by

$$x_\Delta(t + \Delta t) = 2x(t) - x(t - \Delta t) + \Delta^2 M^{-1} F(x(t))$$

To determine the consistency order, we consider the Taylor expansion at $t$ to approximate $x(t+\Delta t)$ and $x(t-\Delta t)$,
\[
(2.133) \quad x(t+Δt) + x(t-Δt) = x(t) + Δt x'(t) + ½ Δt^2 x''(t) + 1/6 Δt^3 x'''(t) + O(Δt^4) + \\
\quad [ x(t) - Δt x'(t) + ½ Δt^2 x''(t) - 1/6 Δt^3 x'''(t) + O(Δt^4) ]
\]

So,

\[
(2.134) \quad x(t+Δt) = 2 x(t) - x(t-Δt) + Δt^2 x''(t) + O(Δt^4)
\]

Now, we subtract \( x_\Delta(t+Δt) \) from the Taylor approximation of \( x(t+Δt) \). Since \( M^{-1} F(x(t)) = x''(t) \), we have

\[
(2.135) \quad x(t+Δt) - x_\Delta(t+Δt) = O(Δt^4)
\]

So, the consistency error for the position is \( O(Δt^4) \). Therefore, \( ε(t,y,Δt) = ||e_\Delta||_\infty = 3 \).

### 2.6.3 Central Verlet algorithm : velocity

The velocity for the central Verlet algorithm is given by

\[
(2.136) \quad v_\Delta(t) = ½ [ x(t+Δt) - x(t-Δt) ]
\]

To determine the local accuracy, we note that

\[
(2.137) \quad x(t+Δt) - x(t-Δt) = x(t) + Δt x'(t) + ½ Δt^2 x''(t) + 1/6 Δt^3 x'''(t) + O(Δt^4) - \\
\quad [ x(t) + Δt x'(t) + ½ Δt^2 x''(t) - 1/6 Δt^3 x'''(t) + O(Δt^4) ]
\]

So,

\[
(2.138) \quad v(t) = ½ [ x(t+Δt) - x(t-Δt) ] / Δt + O(Δt^2)
\]

Now, we subtract \( v_\Delta(t) \) from the Taylor approximation of \( v(t) \):

\[
(2.139) \quad v(t) - v_\Delta(t) = O(Δt^2)
\]

So, the consistency error for the velocity is \( O(Δt^2) \). Therefore, \( ε(t,y,Δt) = ||e_\Delta||_\infty = 1 \).

### 2.6.4 Non-central Verlet algorithm : position

The position for the non-central Verlet algorithm is given by

\[
(2.140) \quad x_\Delta(t+Δt) = 2x(t) - x(t-Δt) + Δt^2 M^{-1} F(x(t-Δt))
\]

Note that the Taylor expansion gives

\[
(2.141) \quad x(t+Δt) = 2 x(t) - x(t-Δt) + Δt^2 M^{-1} F(x(t)) + O(Δt^4)
\]
But,

\[
(2.142) \quad x_{\Delta}(t + \Delta t) = 2x(t) - x(t - \Delta t) + \Delta t^2 M^{-1} \left[ F(x(t)) + \Delta t x''(t) \right] + O(\Delta t^3) = \\
= 2x(t) - x(t - \Delta t) + \Delta t^2 M^{-1} F(x(t)) + \Delta t^3 x'''(t) + O(\Delta t^4) = \\
= 2x(t) - x(t - \Delta t) + \Delta t^2 M^{-1} F(x(t)) + O(\Delta t^4) =
\]

So, the consistency error for the position is \( O(\Delta t^3) \). Therefore, \( \epsilon(t, x, \Delta t) = \|\epsilon_{\Delta}\|_\infty = 2 \).

### 2.6.5 Velocity Verlet algorithm: position

To derive the local accuracy for the velocity Verlet algorithm, we show the equivalence with the central Verlet algorithm for position by computing \( x_{\Delta}(t + 2\Delta t) \) by recursive substitution.

The position and velocity for the velocity Verlet algorithm are given, respectively, by

\[
(2.143) \quad x_{\Delta}(t + \Delta t) = x(t) + \Delta t \cdot v(t) + \frac{1}{2} \Delta t^2 M^{-1} F(x(t))
\]

and

\[
(2.144) \quad v_{\Delta}(t + \Delta t) = v(t) + \frac{1}{2} M^{-1} \Delta t \left[ F(x(t)) + F(x(t + \Delta t)) \right].
\]

Applying (2.143) to advance the solution from \( t + \Delta t \) to \( t + 2\Delta t \), we have

\[
(2.145) \quad x_{\Delta}(t + 2\Delta t) = x_{\Delta}(t + \Delta t) + \Delta t \cdot v_{\Delta}(t + \Delta t) + \frac{1}{2} \Delta t^2 M^{-1} F(x(t + \Delta t)) = \\
x_{\Delta}(t + \Delta t) + \Delta t \cdot v(t) + \frac{1}{2} \Delta t M^{-1} F(x(t)) + \frac{1}{2} \Delta t^2 M^{-1} F(x(t + \Delta t)) = \\
x_{\Delta}(t + \Delta t) + \Delta t \cdot v(t) + \frac{1}{2} \Delta t^2 M^{-1} F(x(t)) + \frac{1}{2} \Delta t^2 M^{-1} F(x(t + \Delta t)) = \\
x_{\Delta}(t + \Delta t) + x_{\Delta}(t + \Delta t) - x(t) + \Delta t^2 M^{-1} F(x(t + \Delta t)) = \\
2 \ x_{\Delta}(t + \Delta t) - x(t) + \Delta t^2 M^{-1} F(x(t + \Delta t))
\]

In line two of (2.145), the equation (2.144) is substituted. In line three of (2.145), the distributive property is applied twice. In line four of (2.145), a substitution is made based on the equation (2.143) and like terms are combined. The end result is exactly the formula for the central Verlet algorithm for \( x_{\Delta}(t + 2\Delta t) \) in terms of two preceding steps and the force evaluated at the previous step. So, the consistency error for the position for velocity Verlet is also \( O(\Delta t^4) \). Therefore, \( \epsilon(t, x, \Delta t) = \|\epsilon_{\Delta}\|_\infty = 3 \).

### 2.6.6 Velocity Verlet algorithm: velocity

The velocity for the velocity Verlet algorithm is given by
\[ v_\Delta(t+\Delta t) = v(t) + \frac{1}{2} M^{-1}\Delta t \left[ F(x(t)) + F(x(t+\Delta t)) \right]. \]

So, we have
\[ (2.147) \]
\[ v(t+\Delta t) = v(t) + \frac{1}{2} M^{-1}\Delta t \left[ F(x(t)) + F(x(t+\Delta t)) \right] = v(t) + \frac{1}{2} M^{-1}\Delta t F(x(t)) + \frac{1}{2} \Delta t M^{-1} F(x(t)) + \frac{1}{2} \Delta t^2 x'''(t) + O(\Delta t^3) \]
\[ v(t) + \Delta t M^{-1} F(x(t)) + \frac{1}{2} \Delta t^2 x'''(t) + O(\Delta t^3) \]

In line two of (2.147), the distributive property is applied. In line three of (2.147), the following Taylor series substitution is made:
\[ (2.148) \]
\[ M^{-1} F(x(t+\Delta t)) = x''(t+\Delta t) = x''(t) + \frac{1}{2} \Delta t x'''(t) + O(\Delta t^3) \]
\[ M^{-1} F(x(t)) + \frac{1}{2} \Delta t x'''(t) + O(\Delta t^3) \]

Note that the Taylor expansion gives
\[ (2.149) \]
\[ v(t+\Delta t) = v(t) + \Delta t M^{-1} F(x(t)) + \frac{1}{2} \Delta t^2 x'''(t) + O(\Delta t^3) \]

Now, we subtract \( v_\Delta(t+\Delta t) \) from the Taylor approximation of \( v(t+\Delta t) \):
\[ (2.150) \]
\[ v(t+\Delta t) - [v_\Delta(t+\Delta t) + O(\Delta t^3)] = O(\Delta t^3) \]

So, the consistency error for the velocity is \( O(\Delta t^3) \). Therefore, \( \epsilon(t, v, \Delta t) = ||v_\Delta||_\infty = 2 \).

### 2.7 References


Figures and Tables

Figure 2.1 Depiction of components of an empirical potential energy function

For each type of component, a ball-and-stick visualization is shown in which atoms are shown as balls and covalent bonds are shown as sticks as well as the contribution to potential energy for a potential energy function of the form described in the text with a typical parameterization. Additional comments on the individual components are shown below:

**Bond Lengths**: Inclusion of springs in ball-and-stick depiction serves to emphasize quadratic form of bond length potential.

**Dihedral Angles**: Angle determined by four atoms. It is the angle between a plane defined to include three atoms and another plane defined to include three other atoms. Two atoms are in both planes. Dihedral angles in which one atom in each plane is bonded to each of the other three, the angle is termed improper. When the four atoms are in sequence in a chain and the two atoms in each plane are the middle two of the chain, the angle is termed proper.

**Electrostatic**: Contribution is negative for oppositely charged atoms; positive for like charged atoms.

**van der Waals**: Repulsive or attractive forces dominate depending on distance between atoms.

This figure was adapted from an unlabeled figure in [Ste2003].
Figure 2.2  Multiple shooting algorithm applied to Mueller potential

Multiple shooting algorithm applied to Mueller potential

The multiple shooting algorithm is applied with four subintervals (i.e. $N = 4$); The time step is $\Delta t = 0.001$;

Beginning and ending time points are $t_0=0$; $t_f=0.29$; Beginning and ending boundary conditions are given by

$$
\begin{align*}
  r( x(t_0,s), v(t_0,s), x(t_f,s), v(t_f,s); s ) &= [ x(t_0,s) - x_0; x(t_f,s) - x_f ]
\end{align*}
$$

where

$$
  x_0 = [-0.5582; 1.4420], x_f = [0.6235; 0.0280].
$$

The Mueller potential energy function is

$$
U(x(t)) = 200e^{(x_1-1)^2-10x_2^2} - 100e^{x_1^2-10(x_2-0.5)^2}
-170e^{-6.5(x_1+0.5)^2+11(x_1+0.5)(x_2-1.5)-6.5(x_2-1.5)^2}
+15e^{-0.7(x_1+1)^2+0.6(x_1+1)(x_2-1)-0.7(x_2-1)^2}
$$

The trajectories for the multiple shooting algorithm are plotted as follows:

Initial guess trajectory -- green.  2nd, 3rd and 4th iterations -- magenta.  5th iteration -- black

A contour plot for $U(x(t))$ shown in the background.  Dark blue regions represent low energy ‘wells’ around two local minima, $x_0$ and $x_f$, of the potential energy surface.
Figure 2.3  Condition numbers for N-acetyl-N'-methylalaninamide in vacuo: C7\textsubscript{eq} \rightarrow C6

Initial coordinates C7\textsubscript{eq}; initial velocity direction C6
\(\kappa(F'(t;s))\) –green; \(\kappa(Y(t;s))\) –blue from IVP simulations

\(\kappa(F'(t;s))\) –green and \(\kappa(Y(t;s))\) –blue from IVP simulations of the alanine dipeptide in vacuo are plotted vs. step number on a logarithmic scale (base 10)

The initial coordinates correspond to the C7\textsubscript{eq} local minimum. The initial velocity direction is toward the C6 local minimum. Energy scaling (left to right; top to bottom) in kcal mol\(^{-1}\) are as follows:

\([-11, -15, -10; -5, 0, +5, +10, +15; +20, +25]\)

The time step is \(\Delta t = 0.015725\) AKMA units \(\approx 0.769\) fs;

10,000 total steps are taken. The total time of simulation is 157.24 AKMA units \(\approx 7.69\) ps.
Figure 2.4  Condition numbers for \( N\text{-}acetyl-N'\text{-}methylalaninamide \) in \textit{vacuo} :  \( C7_{eq} \rightarrow C5_{\beta} \)

Initial coordinates \( C7_{eq} \); initial velocity direction \( C5_{\beta} \)

\( \kappa(F'(t;s)) \) – green; \( \kappa(Y(t;s)) \) – blue from IVP simulations

\[ \kappa(F'(t;s)) \text{ – green and } \kappa(Y(t;s)) \text{ – blue from IVP simulations of the alanine dipeptide in vacuo are plotted vs. step number on a logarithmic scale (base 10).} \]

The initial coordinates correspond to the \( C7_{eq} \) local minimum. The initial velocity direction is toward the \( C5_{\beta} \) local minimum. Energy scaling (left to right; top to bottom) in kcal mol\(^{-1}\) are as follows:

\[ [+1,-15,-10;-5,0,+5,+10,+15,+20,+25] \]

The time step is \( \Delta t = 0.015725 \text{ AKMA units} \approx 0.769 \text{ fs} \); 10,000 total steps are taken. The total time of simulation is 157.24 AKMA units \( \approx 7.69 \text{ ps} \).
Figure 2.5 Condition numbers for N-acetyl-N’-methylalaninamide in vacuo: C7\textsubscript{ax} \rightarrow C7\textsubscript{eq}

Initial coordinates C7\textsubscript{ax}; initial velocity direction C7\textsubscript{eq}
κ(F’(t;s)) –green; κ(Y(t;s)) –blue from IVP simulations

κ(F’(t;s)) –green and κ(Y(t;s)) –blue from IVP simulations of the alanine dipeptide in vacuo are plotted vs. step number on a logarithmic scale (base 10)
The initial coordinates correspond to the C7\textsubscript{ax} local minimum. The initial velocity direction is toward the C7\textsubscript{eq} local minimum. Energy scaling (left to right; top to bottom) in kcal mol\textsuperscript{-1} are as follows:
[+73, −10, −5, 0, +5, +10, +15, +20; +30, +40]
The time step is Δt = 0.015725 AKMA units ≈ 0.769 fs;
10,000 total steps are taken. The total time of simulation is 157.24 AKMA units ≈ 7.69 ps.
For three different IVP simulations for alanine dipeptide in vacuo the 1st step for which the $\kappa(F'(t;s))$ exceeds $\text{TOL}_\kappa$ is recorded. The time step is $\Delta t = 0.015725$ AKMA units $\approx 0.769$ fs. So, 2,000 steps $\approx 31$ AKMA units $\approx 1.538$ ps. And, 10,000 total steps $\approx 157.24$ AKMA units $\approx 7.69$ ps.
Figure 2.7 Transition of a triangle: linear DMI

Evolution of interatomic distances monotonic with uniform relative rate

Transformation of a triangle: linear DMI
Figure 2.8 Transition of a triangle: nonlinear DMI

Evolution of interatomic distances
not (monotonic with uniform relative rate)

Transformation of a triangle: nonlinear DMI
3 A MULTIPLE SHOOTING APPROACH FOR ALL-ATOM SIMULATION OF TRANSITIONS BETWEEN TWO MOLECULAR CONFORMATIONS SPECIFIED BY SETS OF INTERNAL COORDINATES

3.1 Abstract

The transitions of molecules from one conformation to another are an essential part of many biological processes. All-atom molecular dynamics simulation (AA-MDS) is a particular type of molecular modeling in which the motion of the atoms or particles of the molecules of the system are tracked dynamically over a period of time and the motion is governed deterministically by Newtonian equations of motion. AA-MDS is a common way to study molecular dynamics, so it is natural to consider its application to the study of conformational transitions. A boundary value approach to molecular dynamics can be considered when starting and final conformations are known, and conformational transitions between these two conformations are required. The multiple shooting method is a method for obtaining numerical solution to boundary value problems (BVP’s) for ordinary differential equations (ODE’s). The application of the multiple shooting method to BVP’s for ODE’s corresponding to transitions between two molecular conformations specified by sets of internal coordinates is proposed. Strategies and issues related to boundary conditions, assignment of initial parameters, and convergence are described. Results from the study of transitions between local minima of the potential energy surface of an alanine dipeptide are presented. Implications of the methods and results of this work for application of multiple shooting to the study of conformational transitions in larger systems are provided.
3.2 Introduction

This introductory section is divided into two parts: (1) a characterization of protein conformations and (2) a brief description of initial value all-atom molecular dynamics simulation (IV-AA-MDS) and its limitations pertaining to the study of conformational transitions.

3.2.1 Characterization of protein conformations

A protein produced via translation from mRNA, initially exists in an unfolded state. In ‘Thermodynamics of Protein Folding and Stability’ by Alan Cooper, (Chapter 6 of Protein: A Comprehensive Treatise, [Coo1999]), an unfolded state is characterized as

*an ill-defined state, or rather set of states comprising anything that is not recognisably folded. A population of conformations, spanning and sampling wide ranges of conformation space depending on conditions. Usually quite open, irregular, heterogeneous, flexible, dynamic structures - no one molecule is like another, nor like itself from one moment to another...*

For a protein to carry out its biological function, it must transition from the unfolded state to the folded state. This process is called folding. The folded state is characterized in [Coo1999] as

*the biologically active (‘native’) form of the [protein] (usually). Compact, showing extensive average conformational homogeneity with recognisable regions of regular, irregular and motif structures, on a background of dynamic thermal fluctuations...*

Important measurements for assessing the state of a protein are the values of the pairs of dihedral angles around the Cα atoms of the amino acids that make up the peptide chain. These angles are commonly called φ and ψ angles. For an arbitrary amino acid, φ and ψ are defined in a way that is analogous to the way that φ and ψ are defined for the alanine amino acid of N-acetyl-N'-methylalaninamide in section 1.3. Pairwise interatomic distances are also important measurements for assessing the state of a protein. In the unfolded state, the distribution of φ and ψ angles can be quite broad (but not random). By contrast, in a folded state, the φ and ψ angles of many residues will typically populate a specific and narrower range of values.
The characterization in the excerpt from [Coo1999] clearly and accurately gives the impression that the folded state is not a single conformation, but, rather, an collection of conformations with some identifiable features in common. The folded state is a relatively stable state. And, often, the folded state is essentially homogenous as is suggested in this excerpt. But, the stability and homogeneity of the folded state can vary for different proteins. For some proteins, there may be alternate folded states. Some alternate folded states may be detrimental to the cell. In these cases, the alternate folded states may be called a misfolded state. For some proteins, there may also be intermediate states that exhibit some stability. These states are known as metastable states. There always exist equilibria among different folded states, metastable states, unfolded states, and misfolded states. The equilibria can be affected by environmental changes. In many cases, the conformational transitions of molecules between a folded state and another state (e.g. important metastable states, alternate folded states, or misfolded states) can have a vital bearing on important biological processes.

3.2.2 Brief description of initial value all-atom molecular dynamics simulation (IV-AA-MDS) and its limitations

Initial value all-atom molecular dynamics simulation (IV-AA-MDS) can be defined mathematically as an initial value problem (IVP) for ordinary differential equations (ODE's) where the motion of atoms, or particles, in the system are subjected to laws of Newtonian physics. In this context, Newtonian forces on the atoms are calculated as the negation of the gradient of a complex, empirically derived potential energy function which gives the potential energy of the system. For an isolated system, the potential energy is a function of the relative locations of the atoms in the system. Some basic concepts of IV-AA-MDS and potential energy will be discussed in more detail in section 3.3. Initial values can include a stable structure and randomly assigned momenta scaled to satisfy some desired level of total energy in the system. IV-AA-MDS has proven to be a useful means for the study of biomolecules. It can be applied to study local motions near a stable conformation of a molecule and it has also been applied to the study of conformational transitions and folding ([Kim2003], [Zag2001], [Rhe2003]). Loosely defined
but generally accepted theories of protein folding and conformational transitions have emerged (e.g. [Onu2004]) and some of the collective understanding of these processes can be attributed to IV-AA-MDS. But, many questions about the folding and transitioning processes remain and attempts to verify the existing theory may sometimes be anecdotal and/or incomplete.

An IVP for AA-MDS is a mathematical model of the molecular dynamics of a particular system. Like all mathematical models, an IVP for AA-MDS has some limitations. The most important limitations with respect to simulation of folding or conformational transitions of large molecules are computational in nature. Computing an approximate solution to an IVP numerically involves a sequential iterative process of determining a discrete evolution of the system of particles in time on a mesh, or ordered set of time points. To capture the fastest motions which have vibrational frequencies on the order of $10^{-13}$ s, the time steps, or increments between adjacent points of the mesh, are require to be about $10^{-15}$ s. Because of this constraint on the length of the time steps, there are practical limits for the total time interval for a simulation, which are currently on the order of nanoseconds ($1 \text{ ns} = 10^{-9}$ s) or microseconds ($1 \mu s = 10^{-6}$ s), where the range is due to factors such as size of the system (i.e. numbers of particles) to be simulated, computational resources available, and sample size required. In nature, conformational transitions of biological interest can occur, however, over much longer time intervals (e.g. milliseconds ($1 \text{ ms} = 10^{-3}$ s) or even seconds. In summary, there simply is not enough computational resources to simulate many events that would be scientifically interesting to simulate due the combination of the constraint on the length of the time steps, the length of expected waiting times for the events to initiate, the expected duration of the events once they have been initiated. This limitation will be referred to as the time interval limitation. It is worth noting that, in practice, there are also limitations on the number of particles in the system and we will refer to this as the system size limitation. This limitation is a realization that the scope of a molecular dynamics simulation must be limited. It would be easy to identify a seemingly simple system with an excessive number of particles since even a simple one-celled organism is estimated to have hundreds of trillions of atoms.
It is possible that a solution to a specific IVP or BVP may be obtained by analytical means. But, the IVP’s and BVP’s that arise in AA-MDS will, in general, require use of a numerical method implemented on a computer to obtain an approximate solution. The solution is only approximate due to error in the numerical method and error in arithmetic calculations performed in finite machine arithmetic. The term position trajectory, refers to an ordered set of coordinates assumed by a dynamical system on a discrete mesh of time points. The term velocity trajectory refers to an ordered set of velocities corresponding to a position trajectory. A trajectory consists of a position trajectory and a velocity trajectory. It can be directly defined as an ordered set of intermediate states assumed by a dynamical system on a discrete mesh of time points. The $i^{th}$ intermediate state of the trajectory is known as the $i^{th}$ snapshot of the trajectory. For a position or velocity trajectory with $(D+1)$ snapshots including the endpoints, the trajectory will be represented by a $3n \times (D+1)$ matrix. So, columns of this matrix correspond to snapshots of the trajectory. Rows correspond to the evolution of the position or velocity of one particle in one coordinate direction. The relevance of an AA-MDS trajectory is partially limited by the accuracy of the previously described potential energy function. For relevant discussion of issues related to the potential energy function, see [Gar2003], [Pri2002], [Go1983], [Gar2002], [Ued1978], [Ren2006], [Roy2005], [Hu2003]. For our purpose, we do not address this issue, but assume that the potential energy function is adequate for modeling purposes and that AA-MDS can be applied to generate physically meaningful trajectories. For analytical and inferential purposes, an ensemble of trajectories is preferable to a single trajectory. Generally as the ensemble size increases, the utility and reliability of data will increase. The need for sizable ensembles rather than single trajectories further highlights computational limitations of AA-MDS. In light of the important computational limitation, how are and how can interesting, but computationally challenging, folding and transitional events be studied? In the next section, we provide a brief description and characterization of some approaches.
3.3 Preliminaries

3.3.1 Current approaches to the study of conformational transitions of proteins

In this section, we describe current approaches to the study of conformational transitions of proteins given the time interval limitation and system size limitation that were discussed in the previous section. A classification of approaches is introduced in the review paper [Sch1997]. Three different classes of approaches are to generate (1) ordered, accurate trajectories, (2) ordered, approximate trajectories, or (3) unordered samples of configurations—that is, sample configurations which cannot be ordered to form a trajectory, but which may provide useful data to assess important properties of a system. We will use this classification scheme as a basis for some discussion, providing comments when additional classes, or subclasses are suggested. We note that our primary objective is to study trajectories corresponding to conformational transitions, so our discussion of the third class will be limited. Also, we note that it is not our objective to provide a comprehensive review here, but rather just provide some context for our work.

3.3.1.1 Approaches to generate ordered, accurate trajectories

The essential idea behind type (1) approaches limit the problem of study so that it has appropriate combinations of system size, expected waiting time for occurrence, and required duration. IV-AA-MDS is a class (1) method. There are interesting folding events and transitional events for which simulation is computationally feasible. The study of some of these events may be, at least partially, motivated by their use as models for similar events in larger systems and over longer time intervals. So, IV-AA-MDS continues to be used as a tool for these types of folding and transitional studies. Moreover, advances in high-performance computing and the creative use of existing resources (e.g. [Shi2000]) have extended the range of transitional events accessible to IV-AA-MDS.
Two of the aforementioned creative types of IV-AA-MDS simulation are coupled parallel molecular dynamics (CPMD) simulations and replica exchange molecular dynamics (REMD) simulation. Application of these methods can result in expanded computational resources and also more efficient exploration of phase space (see section 2.1 for a definition of phase space). CPMD and REMD simulations are methods of class (1) but also contain elements of sampling methods of class (3). The trajectories that result from either of these methods may include instantaneous discrete changes in velocities microscopically and kinetic energy macroscopically, so that they may not be smooth trajectories. Appropriate analysis of results using these methods should take these observations into account. For further discussion of REMD, see [Zag2001]. For further discussion of CPMD, see [Sug1999], and [Rhe2003].

IV-AA-MDS approaches are desirable in that assignment of initial momenta of the system are typically randomly generated, possibly subject to constraint on the total momentum of the system, so that estimation of transition rates and likelihoods for different reaction pathways is straightforward. In some cases, information about transitions between two known conformations is desired. IV-AA-MDS approaches do not take advantage of information about both the beginning and ending conformations for the desired trajectories. When this information is available, a boundary value approach to all-atom molecular dynamics (BV-AA-MDS) may be appropriate. In [Bol2000] and [Bol2002], a BV-AA-MDS method of class (1) called transition path sampling was introduced. This is a method for rapidly obtaining an ensemble of trajectories corresponding to a transition from one well in a potential energy surface surrounding a local minimum to an adjacent well surrounding another local minimum separated by an energy barrier given that a trajectory which corresponds to the desired conformational transition has already been obtained. Informally, a well in a potential energy surface surrounding a local minimum, or a potential energy well, is a region of $3n$ dimensional space such that if the state of the system (defined by $3n$ atomic coordinates and $3n$ atomic velocities) at some point in time is sufficiently close to the equilibrium point corresponding to the local minimum, the position of the system will always remain measurably close to the local minimum and the kinetic energy of the system will always be bounded. The well surrounding a
named local minima with name ‘A’ will be called the ‘A’ well. See section 2.1 for a more formal definition of a potential energy well and further details. In the context of a complex macromolecule, conformations of different wells would possibly be distinguishable based on some important features of the different conformations such as values of important dihedral angles, existence of specific native contacts, specific hydrophobic interactions, and specific hydrogen bond interactions. In terms of the potential energy surface, a molecule in a transition process may be need to pass through many potential energy wells en route to the ending conformation. The transition path sampling method doesn’t seem to be ideal for problems of this type. For a review of this method, see [Bol2002].

In [Ole1996], a boundary value approach to all-atom molecular dynamics was considered in conjunction with a numerical method, the **stochastic difference equation** (SDE) method. The SDE method was introduced with an emphasis on application to conformational transitions of proteins over long time intervals where IV-AA-MDS is computationally infeasible. This approach, which can be theoretically derived as a direct application of the least action principle ([Gil1992], [Ole1996]) and is partially based on a variational implementation of the Verlet algorithm described in [Gil1992], involves implementation of a finite difference method for simultaneously satisfying boundary value conditions and approximately solving the equations of motion. The term approximately here is used to imply a level of approximation that is significantly less accurate than approximation resulting from numerical solution of an IVP. So, this method was introduced as a class (2) method. But, in [Bai2006], it was shown that this method can be applied with a dense mesh as a BV-AA-MDS method of class (1) to find BV-AA-MDS trajectories (i.e. AA-MDS trajectories that solve BVP’s). In other words, the method can be applied to find trajectories that simultaneously satisfy boundary value conditions and solve the equations of motion (within limits of numerical methods and computer arithmetic).

Another BV-AA-MDS method of class (1) is **multiple shooting**. Multiple shooting (MS) is a numerical method for solving BVP's for ODE's. MS for nonlinear BVP’s is an iterative method. Assuming there are no additional constraints and there are \( n \) particles in the system of interest, then \( 6n \) initial value
problems (IVP’s) are solved on each of \( N \) subintervals of the full time interval of the simulation. The solutions of the IVP’s provide, for each subinterval, a set of \( 3n \) trajectories for the positions of the \( n \) atoms and a set of \( 3n \) trajectories for the velocities of the \( n \) atoms. One application for conformational transitions of proteins can be described as follows. There are \( 6n \) boundary conditions — \( 3n \) for the conformation at the beginning of the simulation and \( 3n \) for the conformation at the end of the simulation. For a solution trajectory, continuity of the \( 3n \) coordinates and \( 3n \) velocities is required at each of \( N–1 \) nodes dividing the subintervals. So, in total, there are \((3n + 3n)(N–1)\)\(+ 3n = 6nN\) conditions. For each of the \( N \) subintervals, there will be \( 3n \) parameters for the initial positions of the \( n \) atoms in the system and \( 3n \) parameters for the initial velocities of the \( n \) atoms, so, in total, there will be a full set of \( 6nN \) parameters. So, to initiate MS for BV-AA-MDS, a \( 6nN \times 1 \) initial parameter vector is required. After the first iteration, typically, the \( 6nN \) position and velocity trajectories will have some jump discontinuities at the \( N–1 \) nodes dividing the subintervals and the \( 6n \) boundary conditions will not all be satisfied. But, a system of \( 6nN \) nonlinear equations (NLE’s) with \( 6nN \) variables (i.e. parameters) can be derived. Applying Newton’s method or some other method for solving systems of equations, each iteration will provide adjustments to the initial positions and initial velocities at the beginning of each of the \( N \) subintervals. On the next iteration, the adjusted initial data, i.e. the adjusted parameter vector, is used to solve another set of \( 6nN \) scalar IVP’s. The solution of the NLE’s will correspond to a trajectory for which positions and velocities are continuous at each node that separates the subintervals and for which the boundary value problem is satisfied. This approach has a natural extension to a parallel or distributed environment since a multiple shooting algorithm can be defined so that the IVP trajectories on the different subintervals can be obtained from separate processors.

In this chapter, we will illustrate the use of a multiple shooting method to solve BVP’s corresponding to transitions between local minima of a potential energy surface in \textit{vacuo} for \( N\text{-acetyl-N’-methylalaninamide} \), which is a capped, or blocked, alanine dipeptide. The restriction of the boundary conditions to two specific sets of internal coordinates (e.g. the two local minima) is convenient mathematically. The number of parameters is equal to the number of equations to satisfy for a solution
trajectory as was described above. Methods for attempting to solve the above-mentioned nonlinear system of equations involve, on each iteration, a linear system solution that can be implemented by a direct approach with complexity $O(Nn^3)$ where $N$ is the number of multiple shooting subintervals and $n$ is the number of atoms in the system. For larger and more complex motions in molecules like proteins, it may be of more practical and greater interest to define boundary conditions corresponding to potential energy wells rather than local minima. Single shooting is a special case of multiple shooting with one subinterval, so $N=1$.

When defining boundary conditions in this way, the number of boundary conditions and the number of parameters can be reduced from $6n$ and $6nN$ respectively as will be described in Chapter 4. Reduction of number of boundary conditions and number of parameters are important for the application of shooting methods to large complex systems. Single shooting methods have been developed to determine molecular dynamics trajectories of small molecules ([Gil1992]) and multiple shooting methods have been applied to determine periodic orbits of small chemical systems ([Far1998]).

### 3.3.1.2 Approaches to generate detailed, ordered, approximate trajectories

Because of the time interval limitation and the system size limitation, it may not always be feasible to find an AA-MDS trajectory between a particular conformation to another particular conformation. The alternative of studying a model problem by AA-MDS may not always be useful. So, it is natural to consider alternatives to AA-MDS. If the premise that AA-MDS is the most detailed and physically realistic simulation or modeling approach is accepted, then any other approach will not be as detailed or not as physically realistic or neither. Collectively, these types of alternative approaches will be called approximate approaches and can be categorized as class (2) using the [Sch1997] categorization.

Based on frequencies of amino acids in protein sequences in the comprehensive catalog of information on proteins UniProt (Universal Protein Resource), the average protein has about 10 hydrogen atoms and about 9 non-hydrogen, or heavy, atoms for an average of about 19 total atoms ([Uni2007]). The level of detail of an approximate approach reflects the number of particles used to model a protein amino
acid. The level of detail will be classified in this dissertation as all-atom or coarse-grained. Sometimes, instead of representing all atoms by a particle, some hydrogen atoms are not explicitly modeled. Rather, a heavy atom and hydrogen atoms bonded to it are modeled as one particle. This approach is called a united atom approach. Evidently, a united atom model will include, on average about nine particles per amino acid. We will not distinguish between all-atom modeling and united atom modeling. Both approaches will be categorized as all-atom approaches. While dividing models into coarse-grained and all-atom involves a somewhat arbitrary ‘dividing line’, for proteins, let’s say that coarse-grained models have no more than an average of five particles per amino acid. We will now provide brief descriptions of some approximate approaches, i.e. class (2) approaches.

An all-atom class (2) alternative approach that was first introduced in [Cze1990] involves an attempt to minimize a discretized version of a line integral representing a path along a potential energy surface between two conformations such as those representing the local minima of two potential energy wells. A feature of the SDE approach that’s emphasized in [Ole1996] and [Elb1999] is the enabling of a much larger time step than is allowed in IVP approaches to AA-MDS. Results presented in [Elb1999] and [Elb2002] suggest that this method is robust in generating trajectories which capture essential features of the AA-MDS trajectories along a common pathway between two different conformations with much of the high-frequency motion filtered out. The SDE method can be applied with time as the independent variable (SDET) ([Elb1996], [Elb1999]) or with length as the independent variable (SDEL) ([Elb2002]). These methods apply the least-action principle and attempt to find a solution to the equations of motion by optimization of a path integral defined using principles of classical mechanics. Computationally, the SDET method is equivalent to application of a finite-difference scheme for BVP’s (e.g. Chapter 5 of [Asc1995]) with a global convergence strategy to minimize the magnitude of the residual (see sections 2.3.3, 2.3.4, 2.3.7, and 2.4.5). Similar approaches have been developed in [Pas2001] and [Gla2004]. For a more detailed review of these methods, see [Elb2004] or [Gla2001].

In [Sch1997], several all-atom class (2) approaches which provide trajectories that appear to closely approximate AA-MDS trajectories were described. These methods allow for relatively small
increases in step size by up to a factor of 5 through the use of constrained dynamics, normal-mode based schemes, Langevin modeling, and multiple time stepping methods. Some methods such as the constrained dynamics algorithm, SHAKE, which allows increase in time step by factor of about 2 with minimal loss of information or accuracy, are widely used in practice.

Conformational transitions in biomolecules have also been studied at the all-atom level with normal mode analysis (NMA) (e.g. [Bro1983], [Tam2001], [Jaa1998]. NMA classifies the possible deformations of a protein by their energetic cost. Collective motions are energetically cheaper than local ones ([Hin2004]). NMA is theoretically applicable only to oscillatory motions around a local minimum. In spite of this narrow range of applicability in theory, NMA-based methods for finding feasible pathways between two structures have been shown to be useful in practice ([Tam2001], [Kun2004], [Son2006]). These methods, in general, involve following low-frequency modes and calculating normal modes as they move along a path. A straight-forward application of this method can become computationally expensive when the number of atoms, $n$, in the system grows large as the approach can require inversion of $3n$ by $3n$ matrix. This makes straight-forward application of normal-mode analysis an $O(n^3)$ method. However, modified normal-mode methods have been developed in which the incremental computational cost grows on the order of only $O(n)$.

Distance matrix interpolation (DMI) is another approach to generate detailed, ordered, approximate trajectories. DMI was first introduced as a coarse-grained approach, and the coarse-grained approach is similar to the all-atom approach. Discussion of DMI included in the next section.

3.3.1.3 Approaches to generate coarse-grained, ordered, approximate trajectories

In terms of level of detail, as an example, in a coarse-grained approach, the smallest unit might represent one residue or perhaps a selected number of residues, or perhaps each residue of a protein could be identified by two points — the location of the C$\alpha$ atom and some other point that represents the side chain. An advantage of the coarse-grained methods is that they can be applied to very large systems without significant computational cost. The Gerstein lab at Yale uses a coarse-grained representation of
biomolecules (e.g. one point per residue). For molecules with two different conformations, this lab has
developed a computationally efficient method for rapidly creating an sequence of intermediate
conformations between the two conformations using Cartesian interpolation and constrained energy
minimization ([Ech2003], [Ger2004]).

In distance matrix interpolation (DMI) approaches, a trajectory with a specified number of
intermediate conformations is desired. For each desired intermediate conformation, a matrix of target
distances and an objective function is defined. The objective function is defined as a search for a set of
internal coordinates for the molecule for which the overall differences between a matrix of interatomic
distances and the matrix of target distances is minimized (in a way that will be described in Chapter 5).
The DMI approach was introduced as a coarse-grained approach that involved an elastic network model
(DMI-ENM). With minor modifications, it can be applied at the all-atom level of detail as well. In DMI-
ENM, the previously mentioned objective function leads is approximated by a quadratic function which
leads to a quadratic model for which a solution can easily be generated. Some studies ([Kim2003]) suggest
that the series of DMI-ENM intermediate conformations provide a good approximation to transition
trajectories generated from all-atom simulations. For further discussion of DMI-ENM, see [Kim2002a] or
[Kim2002b] and Chapter 5 of this dissertation. DMI methods that use an objective function that is often
used in the field of distance geometry are introduced in Chapter 5 of this dissertation. These methods have
shown promise as a useful method for generating DMI trajectories. (In this dissertation, we call the
trajectories generated by this method DMI-DG trajectories.) A common and intuitive way to generate the
matrices of target distances is by independently interpolation for each matrix entry. For each entry,
interpolation by a linear polynomial is performed between the values in the distance matrix entries for the
beginning and ending conformations. As an alternative, spline interpolation is introduced as a way to
generate target distance matrices in Chapter 5. Additionally in that chapter, the refinement of DMI
trajectories by the use of local constrained energy minimization (LCEM) is described. An all-atom DMI
trajectory can be used for generating initial trajectories required in the use of the multiple shooting method
as a numerical method to solve BVP’s for AA-MDS.
In [Kol2004] a coarse grained lattice model is described in which $C^\alpha$ carbons are connected by pseudo-atoms and side chains are represented by 1 to 2 points. An REMD simulation was performed using this model in [Mal2005] in which misfolding of some prion-like sequences was observed in the presence of a misfolded protein. The force field was determined strictly using statistical data from the PDB and water molecules were not explicitly included. This method allows for simulations longer than IV-AA-MDS by approximately two orders of magnitude ([Kol2005]).

3.3.2 All-atom molecular dynamics simulation (AA-MDS)

All-atom molecular dynamics simulation (AA-MDS) generally refers to a particular type of molecular modeling in which the motion of the atoms or particles of the molecules of the system are tracked dynamically over a period of time and the motion is governed deterministically by the Newtonian equations of motion. More specifically,

\[ M \ddot{x}(t) = f(x(t)), \quad t_0 < t < t_f \]

where $t$ is a scalar representing time where $t_0 < t < t_f$; $x(t)$, $v(t)$, and $a(t)$ are $3n \times 1$ vectors representing the position, velocity, and acceleration, respectively, of the $n$ particles of the system at time $t$ in three dimensions of a rectangular coordinate system; $M$ is a $3n \times 3n$ diagonal matrix with the mass of each particle repeated in the three diagonal entries corresponding to the three dimensions of physical space; and $f(x(t))$ is an $3n \times 1$ vector representing the force acting on each particle of the system at time $t$ in each dimension.

Note that $v(t) = x'(t)$ and $a(t) = x''(t)$, so (3.1) is a 2\textsuperscript{nd}-order ordinary differential equation (ODE). In order to begin a simulation, additional specifications are required.

3.3.3 Initial value AA-MDS (IV-AA-MDS)

For IV-AA-MDS, additional specifications are the initial values of the form

\[ x(t_0) = \dot{x}, \quad v(t_0) = \dot{v} \]

where $\dot{x}$ and $\dot{v}$ are $3n \times 1$ vectors and $t_0 \leq t \leq t_f$. Equations (3.1) and (3.2) define an initial value problem (IVP). For $f$ linear, the domain for existence and uniqueness of solutions can be specified by inspection of
3.3.4 Boundary value AA-MDS (BV- AA-MDS)

For a two-point boundary value (BV-) AA-MDS, additional specifications are given by

\[(3.3) \quad r \left( x(t_0), v(t_0), x(t_f), v(t_f) \right) = 0 \]

where \( r \) is an \( R \times 1 \) vector for some integer \( R \). In this chapter, we assume \( R = 6n \) and that the \( 6n \) boundary conditions can be separated into \( 3n \) conditions on \( x(t_0) \) and \( 3n \) conditions on \( x(t_f) \). So, the number of scalar boundary conditions is equal to the number of scalar differential equations in (3.1). In Chapter 4, we will consider BVP’s with \( R < 6n \). Equations (3.1) and (3.3) define a two-point boundary value problem (BVP).

For a BVP of this form, in general, there may or may not be a solution, and if there is one solution, it might not be the only one. The adjective ‘two-point’ indicates that \( r \) is a function describing the characteristics of the system at two time-points, \( t_0 \) and \( t_f \). We assume \( f(x) \) is the gradient of a real-valued function of \( x \) called a potential energy function, \( U(x) \). So, \( f(x) = -\nabla U(x) \).

3.3.5 Multiple shooting (MS)

A brief, informal description of the numerical method for solving BVP’s that is known as multiple shooting was given in section 3.3.1.1. For a more formal description of multiple shooting methods as they apply to BV-AA-MDS, it is convenient to rewrite the system of \( 3n \) equations in the \( 2^{nd} \)-order ODE (3.1) equivalently as a system of \( 6n \) equations that define a \( 1^{st} \)-order ODE as shown in equation (3.4) below.

Then, the BVP (3.1), (3.3) can be written in the form below

\[(3.4) \quad y'(t) = h(y(t)), \quad t_0 < t < t_f \]

\[(3.5) \quad r(y(t_0), y(t_f)) = 0 \]

where \( r(y(t_0), y(t_f)) = r(x(t_0), v(t_0), x(t_f), v(t_f)) \). Since we are assuming \( R = 6n \) in this chapter, \( r(y(t_0), y(t_f)) \) is a function with \( 6n \) components.
For numerical solution of an ODE on an interval \([t_0, t_f]\), a mesh is a set of discrete points contained within the interval \([t_0, t_f]\). Continuing with the description of multiple shooting methods, following [Asc1995], we subdivide \([t_0, t_f]\) into \(N\) subintervals using a mesh \(\Delta^s: \{t_i: 0 \leq i \leq N\}\) such that 
\(t_0 < t_1 < \ldots < t_N = t_f\). Then, we will solve IVP’s on each subinterval:

\[
(3.6) \quad y'(t; s) = h(y(t; s)), \quad y(t_i) = s_i, \quad t_i < t < t_{i+1}, \quad 0 \leq i \leq N-1.
\]

where \(s = [s_0; s_1; \ldots; s_{N-1}]\) is a parameter vector in which each block component, \(s_i\), contains initial conditions at \(t = t_i\). The notation with the semicolon followed by \(s\) provides a reminder that the solution is dependent on the initial conditions which are determined by \(s\). There are \(6nN\) unknown parameters. The solution for a given \(s\) is

\[
(3.7) \quad y(t) \equiv y(t; s_i), \quad t_i < t < t_{i+1}, \quad 0 \leq i \leq N-1.
\]

We want to find \(s^*\) such that

\[
(3.8) \quad y_i(t_{i+1}; s_i^*) = s_{i+1}, \quad 0 \leq i \leq N-1
\]

\[
(3.9) \quad r(s_1^*, y_N(t_f; s_{N-1}^*)) = 0.
\]

We define

\[
(3.10) \quad F(s) \equiv \begin{bmatrix}
    y_0(t_1; s_0) - s_1 \\
    y_1(t_2; s_1) - s_2 \\
    \ldots \\
    y_{N-2}(t_{N-1}; s_{N-2}) - s_{N-1} \\
    y_N(t_f; s_{N-1}) - y_N(t_f; s_{N-1}) \\
    r(s_0, y_N(t_f; s_{N-1}))
\end{bmatrix},
\]

and want to find solutions of

\[
(3.11) \quad F(s) = 0.
\]

Various global convergence schemes can be used. For the applications to be described in this work, we have used two different iterative global convergence schemes for the above equation—(1) a dogleg trust region algorithm with residual reduction criterion and (2) a damped Newton algorithm with natural
monotonicity reduction criterion. For any iterative approach with iterates given by \( s^0, s^1, \ldots, s^k \), we may write

\[
(3.12) \quad s^{k+1} = s^k + \xi^k.
\]

It is expected that, using either of the two global convergence schemes indicated above, \( \xi^k = \xi^{N_{\text{wtn}}(k)} \) in the final steps of a converging sequence where \( \xi^{N_{\text{wtn}}(k)} \) is the Newton step on the \( k^{th} \) iteration. The Newton step is generated by solving

\[
(3.13) \quad F'(s^k)\xi^{N_{\text{wtn}}(k)} = -F(s^k)
\]

where \( F'(s) \) is the Jacobian of \( F(s) \). If \( F' \) is nonsingular, then

\[
(3.14) \quad \xi^{N_{\text{wtn}}(k)} = -F'(s^k)^{-1} F(s^k).
\]

Here,

\[
(3.15) \quad F'(s) =
\begin{bmatrix}
Y_0(t_1) & -I & 0 & \cdots & \cdots & 0 \\
0 & Y_1(t_2) & -I & \ddots & \ddots & \vdots \\
\vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\
\vdots & \ddots & \ddots & \ddots & \ddots & 0 \\
0 & \ddots & \ddots & 0 & Y_{N-2}(t_{N-1}) & -I \\
B_0 & 0 & \cdots & \cdots & 0 & B_f Y_{N-1}(t_f)
\end{bmatrix}
\]

where

\[
(3.16) \quad Y_i(t) = \frac{\partial y_i(t; s_i)}{\partial s}, \quad 0 \leq i \leq N-1
\]

and

\[
(3.17) \quad B_0 = \frac{\partial}{\partial u} r(u,v), \quad B_f = \frac{\partial}{\partial v} r(u,v).
\]

at \( u=y(t_0; s)=s \) and \( v=y(t_f; s) \). For each \( i \), applying Theorem 7.1.8 of [Sto2002], we can find each \( Y_i \) numerically, step-by-step as we solve equation (3.7) by solving the following matrix ODE:

\[
(3.18) \quad \frac{d}{dt} Y_i(t; s_i) = \frac{\partial}{\partial y} h(y_i(t; s_i)) Y_i(t; s_i), \quad t_i < t < t_{i+1}
\]
In general, the matrices, \( Y_i(t_i) \), will be dense despite the structural sparseness of \( \partial h/\partial y \).

Due to intrinsic error in the numerical methods for solving ODE’s and also because of numerical errors associated with finite arithmetic, the best that can be achieved is an approximate solution to the BVP. The accuracy of the approximation as a Newtonian trajectory and as a solution to the BVP can be expected to depend on the accuracy of the method for numerical solution of IVP’s, mesh selection for IVP solutions, MS mesh selection, tolerance selection for determination of a numerical solution to \( F(s) = 0 \).

### 3.3.6 Single shooting (SS)

Multiple shooting applied with one subinterval, that is with \( N = 1 \), is often referred to as single shooting. In the case of single shooting, we have \( F(s) \)

\[
F(s) \equiv r(s_0, y_0(t_f; s_0)).
\]

and the Jacobian, \( F'(s) \), takes the slightly different form

\[
F'(s) = B_0 + B_f Y_0(t)
\]

### 3.3.7 Numerical solution of initial value problems (IVP’s)

In the shooting methods for the solution of the BVP’s, there are many IVP’s to be solved (see equations (3.7) and (3.18)). In this section, we discuss the numerical solution of these IVP’s. In [Ver1967], L. Verlet applied a discretization scheme for the solution of AA-MDS IVP’s of the form (3.1), (3.2) that was originally proposed by C. Stormer ([Sto1907]) and is commonly known as the Verlet algorithm. The Verlet algorithm is a widely used and well-known algorithm. The time step for the Verlet algorithm is constant. Let \( \Delta t \) be the time step. Using subscripts to denote iteration number, this algorithm is given by

\[
x_{i+1} = 2x_i - x_{i-1} + \Delta t^2 M^{-1} f(x_i)
\]

\[
v_i = (x_{i+1} - x_{i-1})/(2\Delta t)
\]
This algorithm is easily derived, easy to implement, efficient, and has a desirable property known as the 
*symplectic* property. For more on this property, see [Lei2004] or [Hai1993]. The Hamiltonian or total 
energy function is a measurement of total energy and is given by

\[
H(x, m, v) = E_{\text{TOTAL}}(x, m, v) = U(x(t)) + \frac{1}{2} \sum_{i=1}^{n} m_i \left( v_{3i-2}(t)^2 + v_{3i-1}(t)^2 + v_{3i}(t)^2 \right)
\]

where \( m \) is an \( n \times 1 \) vector of atomic masses. For Newtonian mechanics with a fixed number of particles 
and fixed volume, \( H \) is theoretically constant as a function of time. This implies that theoretically total 
energy is conserved. A trajectory generated by a numerical method will not, in general, exhibit 
conservation of energy in precise analytical terms, but it should exhibit an conservation of energy in an 
approximate sense. The capacity to maintain an approximately constant value of the Hamiltonian is an 
important attribute for a numerical method. The constant value of \( H \) effectively limits the range of 
accessible phase space (see section 2.4.2). Some numerical methods fare better than others in this regard. 
In comparison with non-symplectic algorithms or variable-step-size algorithms, the Hamiltonian is well-
maintained when symplectic schemes like the Verlet algorithm are implemented with constant step size. 
L. Verlet, J.M. Sanz-Serna, and S. Reich all contributed to the understanding of this phenomenon 
([Deu2002]).

Suppose that we want to solve an IVP of the form (3.1), (3.2) on \([t_0, t_f]\). Note that (3.22) does not 
specify how to compute \( x_1 = x(t_0 + \Delta t) \). In IV-AA-MDS, this can be handled by different ways, but, as long 
as the approach is reasonable, it would be expected to have an insignificant effect on results. One can 
estimate \( x_1 \) using the values of \( x_0 \) and \( v_0 \) or use a different method for the 1st step that requires \( x_0 \) and \( v_0 \) but 
not \( x_{-1} \). Neither of these approaches is desirable in the numerical solution of IVP’s for MS. The former 
approach stipulates that the parameter vectors \( s_i \) in (3.6) give conditions for the system at \( t=t_i \) and at 
\( t=t_i - \Delta t \). This does not fit the standard MS model in which each \( s_i \) gives conditions for the system at \( t=t_i \). 
(While a modified MS model could be developed, we will see later that there is another approach that is 
preferable.)
We claim that using a different method for the 1st step approach is also inconvenient for MS. An explanation follows: Suppose that an approximate solution to an IVP of the form (3.1), (3.2) is generated using the Verlet algorithm for a mesh $\Delta t$ on $[t_0, t_f]$. Now, suppose $[t_0, t_f]$ is divided into $N$ non-overlapping subintervals such that each node that divides the subintervals is an element of the $\Delta t$. Consider the set of IVP’s consisting of initial conditions for each subinterval taken to be the position and velocity of the solution to the IVP on $[t_0, t_f]$. One can then concatenate the resulting trajectories on the subintervals into one trajectory on $[t_0, t_f]$. This trajectory will, in general, be a different trajectory than the initial solution on $[t_0, t_f]$. This suggests, in general, that this approach applied to MS, will give trajectories that are impacted in a spurious way by the selection of subintervals, due to properties of the IVP algorithm, not the BVP being solved. This type of inconsistency is not desirable, in general, as it complicates validation of methods and blurs the relationship between an approximate solution to a BVP generated by MS and a solution to that same BVP.

Fortunately, there is another algorithm, the velocity Verlet algorithm, that retains the appealing properties of the Verlet algorithm, including the symplectic property and features a discrete evolution operator that only require the information about the discrete evolution at the previous mesh point. Also, this algorithm has a higher overall convergence order than the Verlet algorithm. The convergence order for the position using the Verlet algorithm is three (i.e. local error of the positions is $O(\Delta t^4)$), but the convergence order for the velocity is only one (i.e. local error for the velocities is only $O(\Delta t^2)$). For the multiple shooting methods, the accuracy is important in calculation of both position and velocity. But, the overall convergence order of the Verlet algorithm is only one. The velocity Verlet algorithm provides an improvement in accuracy of velocity estimates; the convergence order for velocity is two. We apply the velocity Verlet algorithm with constant time step both for the IVP’s in (3.7) and for the solution of the variational equation (3.18) to compute the sensitivity matrices, $\{Y_i\}$. Again, using subscripts to denote iteration number, the velocity Verlet algorithm is given by

$$(3.25) \quad x_{i+1} = x_i + \Delta t \ v_i + \frac{\Delta t^2}{2} f(x_i)$$
\[ v_{i+1} = v_i + \frac{\Delta t}{2} \left( f(x_i) + f(x_{i+1}) \right) \]  

(3.26)

(See the appendix of Chapter 2, which is labeled as section 2.6 for verification of the statements about convergence order in this section.)

One way to classify numerical methods for solving IVP’s for ODE’s is by the number of previous steps required to advance the solution. For a numerical solution \( y_{\Delta} \) on a uniform mesh with constant time step, \( \Delta t \), a one-step method requires only the current value of \( y_{\Delta} \), i.e. \( y_{\Delta}(t) \), (and \( h(y_{\Delta}(t)) \)) to determine \( y_{\Delta}(t+\Delta t) \). A multistep method requires the current value, \( y_{\Delta}(t) \), and previous values. For example a two-step method (with a constant time step requires \( y_{\Delta}(t-\Delta t) \) and \( y_{\Delta}(t) \) to determine \( y_{\Delta}(t+\Delta t) \). The velocity Verlet algorithm is a one step method (although it can be written as a two step method with the step size cut in half as described in [Sch2002]) whereas the central Verlet algorithm is a two step method. It is worthwhile to point out that the arguments for the velocity Verlet algorithm and against the central Verlet algorithm could probably be generalized. So, for the MS methods used in the AA-MDS applications of this dissertation, the arguments above can be considered to be arguments for the use of one-step methods and against the use of multistep methods. This generalization could be investigated in the future.

### 3.3.8 Global convergence algorithms

Two Newton-like global convergence algorithms have been implemented. There is a dogleg trust-region optimization algorithm, based on descriptions from section 6.5 of [Den1996] and chapter 4 of [Noc2002]. The objective function defined to be the mass-weighted magnitude of the residual vector (the vector function \( F \) in (3.10)). A reduction criterion is incorporated into this method that is intended to insure sufficient decrease in the magnitude of the objective function. Also, there is a damped Newton algorithm, based on a description in [Asc1995], which also includes a reduction criterion. This criterion compares magnitude of the correction factor of the current Newton step with magnitude of a hypothetical correction factor using the Jacobian matrix, (the matrix function \( F' \) in (3.15)) of the current step and the residual vector of the proposed next step. This reduction criterion is known as the natural monotonicity criterion, or natural monotonicity test. For both methods, the sequence of final steps of a converging sequence is
expected to be a sequence of Newton steps for the nonlinear system of equations. Both of these global convergence algorithms are described in more detail in section 2.3.4 of this dissertation and more generally in [Deu2004].

3.4 Ideas, methods, and analysis

3.4.1 Specification of boundary conditions for AA-MDS

Boundary conditions can be fairly easily defined in terms of absolute position in three dimensional space. A conformation of a molecule is determined by relative locations of atoms, or equivalently, by internal coordinates. A solution to a BVP with boundary conditions defined in terms of absolute position requires, then, not only that the molecule changes from one conformation to another over the specified time interval, but also that the molecule translates and rotates in space in the way that is specified by the boundary conditions. Boundary conditions defined this way are linear. This leads to a form of BVP with in which \( r(x(t_0), x(t_f)) \) is defined as

\[
\begin{align*}
(3.27) \quad r(x(t_0), x(t_f)) &= [x(t_0) - x_0; x(t_f) - x_f] \\
\end{align*}
\]

where \( x_0 \) and \( x_f \) correspond to rectangular coordinates of the desired structures of the molecule at \( t = t_0 \), and \( t = t_f \), respectively.

From an analytical perspective the absolute boundary conditions might seem unnecessarily restrictive. In the study of conformation transitions, the change in relative location of atoms is of primary importance. Translational or rotational motion is expected to be irrelevant for most purposes. Boundary conditions may be defined so that the absolute locations of a molecule at the endpoints are not specified, but the conformation of the molecule does meet specific criteria that essentially determine the internal coordinates of the molecule. On each iteration, the internal coordinates of the desired final structure are projected (by optimal rotation and translation of the molecule) onto the location of the molecule at the end of the simulation. This leads to a form of BVP with nonlinear boundary conditions in which \( r(x(t_0), x(t_f)) \) is defined as
To begin the explanation of the terms $ALIGN(m,x(t_0),x_0)$ and $ALIGN(m,x(t_f),x_f)$, let $x_0$ and $x_f$ represent rectangular coordinates of the molecule in the conformations that are desired at the beginning and the end of the conformational transition, respectively. The absolute location of $x_0$ and $x_f$ in the rectangular coordinate system is not important. Let $\zeta$ be an arbitrary $3n \times 1$ coordinate vector. We can define a function that transforms $\zeta$ into an $n \times 3$ representation of the coordinates. For notational convenience, let $\tilde{\zeta}$ represent this function applied to $\zeta$. Similarly, define $\eta(\tilde{\zeta})$ to be a function that transforms $\tilde{\zeta}$ from an $n \times 3$ representation of the coordinates into a $3n \times 1$ representation. Now, let $m$ be an $n \times 1$ vector of the atomic masses of the atoms of the system, let $\bar{x}^{(m)}$ be an $n \times 3$ matrix with the mass-weighted mean of the conformation repeated in each row, and let $R^{\min}(m, \tilde{x}, \tilde{y})$ be the $3 \times 3$ rotation matrix that optimally rotates the conformation of an $n \times 3$ matrix, $\tilde{y}$, onto the conformation of an $n \times 3$ matrix, $\tilde{x}$ with respect to the mass vector, $m$. This optimal rotation matrix can be found using singular value decomposition as described in section 12.4 of [Gol1996]. Let the notation $\|\cdot\|_F$ be used to represent the Frobenius norm, and let $\tilde{R}$ be an arbitrary rotation matrix. The Frobenius norm of an arbitrary matrix $A = \{a_{ij}\}$ can be computed using the formula $\|A\|_F = (\sum_i \sum_j |a_{ij}|^2)^{1/2}$. We may write
\[
(3.29) \quad R^{\min}(m, \tilde{x}, \tilde{y}) = \arg \min_R \|m^T (\tilde{x} - \tilde{y}\tilde{R})\|_F^2,
\]
A preferred method for finding $R^{\min}(w, \tilde{x}, \tilde{y})$ uses quaternions ([Cou2004]). In contrast to the method of [Gol1996], this method features a convenient way to exclude orthogonal transformation matrices that contain reflections. Now, define
\[
(3.30) \quad ALIGN(m,x,y) = \eta(\bar{x}^{(m)} + (\tilde{y} - \bar{y}^{(m)}) R^{\min}(m, \tilde{x} - \bar{x}^{(m)}, \tilde{y} - \bar{y}^{(m)}) )
\]
So, the desired boundary conformations are optimally translated and rotated onto the locations of the molecule at $t = t_0$, and $t = t_f$ using the following formulas
\[
(3.31) \quad ALIGN(m,x(t_0),x_0) = \eta(\bar{x}^{(m)}(t_0) + (\tilde{x}_0 - \bar{x}_0^{(m)}) R^{\min}(m, \tilde{x}(t_0) - \bar{x}^{(m)}(t_0), \tilde{x}_0 - \bar{x}_0^{(m)}) )
\]
\[
(3.32) \quad ALIGN(m,x(t_f),x_f) = \eta(\bar{x}^{(m)}(t_f) + (\tilde{x}_f - \bar{x}_f^{(m)}) R^{\min}(m, \tilde{x}(t_f) - \bar{x}^{(m)}(t_f), \tilde{x}_f - \bar{x}_f^{(m)}) )
\]
Note that the method above would be functional for an arbitrary weight vector of non-negative real components. The choice of a weight vector of atomic masses is appealing since the center of mass of \( \text{ALIGN}(m, x(t_0), x_0) \) and \( \text{ALIGN}(m, x(t_i), x_i) \), have the same center of mass as \( x_0 \) and \( x_p \), respectively. Finally, note that a common measure of the similarity of conformation \( x \) and conformation \( y \) is given by the mass-weighted root mean squared deviation (RMSD) (after optimal rotation and translation) which can be computed as

\[
\text{RMSD}(m, x, y) = \frac{\| x - \text{ALIGN}(m, x, y) \|_2}{\sqrt{n}}
\]

### 3.4.2 N-acetyl-N'-methylalaninamide - a blocked alanine dipeptide

\( N\text{-acetyl-N'}\text{-methylalaninamide} \) is a twenty-two atom molecule consisting of an \( \alpha \)-amino acid with an alanine side chain and capped with an acetyl group at the N-terminus and amide and methyl groups at the C-terminus (Figure 3.1). It is commonly called an alanine dipeptide. In this chapter, we study this molecule in \textit{vacuo}, which means that the system doesn’t include any solvent molecules or other environmental molecules. The potential energy of the system is related to Newton’s equations of motion by

\[
f(x(t)) = -\nabla U(x(t)).
\]

The potential energy function is our MATLAB implementation of the AMBER99 potential energy function. The AMBER99 potential energy function takes the same form as the AMBER94 force field which is described in [Cor1995], but includes revised parameter values ([Cas2004]). The form of this potential energy function is given by:

\[
U(x) = \sum_{\text{bonds}} K_b (b - b_0)^2 + \sum_{\text{bond angles}} K_\theta (\theta - \theta_0)^2 + \sum_{\text{dihedral angles}} K_\phi \cos(n\phi + \delta) + \sum_{\text{non-bonded}} \frac{A_{i,j}}{r_{ij}^{12}} - \frac{B_{i,j}}{r_{ij}^6} + \sum_{\text{contacts}} \frac{q_i q_j}{\rho_{ij}} + \sum_{\text{contacts}} \frac{A_{i,j}}{r_{ij}^{12}} - \frac{B_{i,j}}{r_{ij}^6} + \sum_{\text{contacts}} \frac{q_i q_j}{\rho_{ij}}
\]
where $K_b$ is the bond stretching force constant; $b$, the bond length; $b_0$, the ideal bond length; $K_\theta$, the bond angle bending force constant; $\theta$, the bond angle; $\theta_0$, the ideal bond length; $K_\Phi$, the dihedral angle bending force constant; $\Phi$, the dihedral angle; $n$, the dihedral angle multiplicity term; $\Delta$, the phase factor; $A_{i,j}$, the van der Waals repulsion parameters; $B_{i,j}$, the van der Waals attraction parameters; $r_{i,j}$, interatomic distances; $q_i$, atomic electrostatic charges; $\varepsilon$, the dielectric constant; $v_{1-4}$, the van der Waals adjustment factor for 1-4 atom pairs; $l_{1-4}$, the electrostatic adjustment factor for 1-4 atom pairs. A similar potential energy function is implemented in the software MOIL ([Elb1994]). Introductory treatment of potential energy functions for biomolecular dynamics simulation can be found in chapter 8 of [Sch2002]. A brief introduction is also provided in the appendix of Chapter 2 of this dissertation. For all the simulations described in this document, the dielectric constant is set to correspond to a vacuum environment. This is accomplished by setting $\varepsilon = 1$.

Conformational transitions of this alanine dipeptide have been studied previously with simulations in vacuum and in solution (e.g. [Ole1996], [Bol2000], [Flo1969], [Wei1986], [Cor1995], [Hu2004], [Che2004]) and via experiment ([Wei1986], [Cor1995], [Flo1969]). Two measurements—the C-N-Cα-C dihedral angle ($\phi$) and the N-Cα-C-N dihedral angle ($\psi$)—are termed ‘soft’ degrees of freedom for this molecule and are of primary importance in determining the overall shape of the alanine dipeptide. These dihedral angles are identified in Figure 1.2 of Chapter 0 of this dissertation. For the temperature and environments of interest, most of the other internal degrees of freedom deviate only slightly from mean values as a function of time or are not influential in determining the overall shape. Because of the relative flexibility and the importance of the $\phi$ and $\psi$ dihedral angles, it is common to use a projection onto a two-dimensional subspace determined by the values of $\phi$ and $\psi$ to visualize the potential energy surface and also as a way to visualize conformational changes. A two dimensional adiabatic energy map for these two angles, constructed in MATLAB by constrained potential energy minimization, is shown in Figure 3.2. This energy map was constructed as follows: 1369 different fixed combinations of evenly spaced values of $\phi$ and $\psi$ separated by 10º on the $\phi$-$\psi$ plane were specified. The constrained minimization was realized by adding penalty terms to the potential energy function for deviations of $\phi$ and $\psi$ from the fixed values and
then calling MATLAB’s built-in unconstrained minimization function, \texttt{fminunc}). All of the 1369 MATLAB \texttt{fminunc} function calls terminated successfully and about 95\% of them terminated with one of the combinations of $\phi$ and $\psi$ values identified in Figure 3.2 and described in Table 3.1. The contour plot of Figure 3.2 is similar to contour plots from [Ole1996] and [Bol2000] which were produced using similar methods for similar potential energy functions. This figure includes the identification of the $\phi$ and $\psi$ values for six common local minima. This energy map suggests that at least four of the local minima of the potential energy surface — $C7_{eq}$, $C6$, $C5_{\beta}$, and $C7_{ax}$ — represent minima on this adiabatic energy map. The four most common conformations upon termination of the MATLAB function calls are depicted in Figure 3.3.

With respect to the location of local minima on the adiabatic potential energy map, there are some variations between different studies. In a recent publication by Chekmarev, Ishida, and Levy on conformational transitions of alanine dipeptide ([Che2004]), it is asserted that the local minima are typically found in five primary regions—$C7_{eq}$, $C5_{\beta}$, $C7_{ax}$, $\alpha_{R}$, and $\alpha_{L}$ (see Figure 3.2) of the $\phi$-$\psi$ space. With respect to the alanine dipeptide potential energy surface and the frequencies of the different conformations upon termination of the MATLAB function calls, $\alpha_{L}$ and lower $\alpha_{R}$ regions seem to have higher energies and lower frequencies than suggested by studies of the alanine dipeptide in solution. This may be due to the fact that our energy surface corresponds to alanine dipeptide in a vacuum rather than in solution. Also, it could be partly due to the peculiarities of the optimization methods used to model the potential energy surface. Similar explanations may explain the existence of the $C6$ local minimum on this energy surface. For the purpose of testing the multiple shooting methods on the alanine dipeptide, these differences do not seem to be critical.

### 3.4.3 Toward a parameter selection strategy

Specification of boundary conditions, the topic of subsection 3.4.1, is an important step toward the application of a numerical method to BV-AA-MDS. In this subsection, we consider some other important steps. The differential equation is autonomous, so the starting time is arbitrary. We will always set $t_0 = 0$. 
So, the selection of the ending time, $t_f$, of the time interval is equivalent to selection of the total time of the simulation. Also, the $N-1$ nodes which divide the MS subintervals along with the beginning and ending times, $t_0$ and $t_f$, collectively form an $(N+1) \times 1$ MS mesh vector, or simply a mesh, and the *initial parameter vector*, $s^0$, need to be specified. For the BVP’s of this chapter, $s^0$ is $6nN \times 1$ vector consisting of a full set of atomic positions and atomic velocities at each of the $1^{\text{st}} N$ nodes of the mesh. The *initial parameter vector*, $s^0$, may also be called the *initial parameter set*, or simply the *initial parameters*. In practice, there is usually a limited range of values for initial parameters that result in eventual convergence. The ending time, $t_f$, and the mesh selection can also impact convergence. We will refer to the ending time, the mesh, and the initial parameter vector collectively as the *augmented initial parameter set* or the *augmented initial parameters*. Selection of the augmented initial parameters is important. In subsection 3.4.4, we will outline an overall strategy to generate a set of augmented initial parameters, apply the MS algorithm that was described in section 3.3.5, and, if necessary, refine parameters and reapply the MS algorithm.

Let us call the trajectory that results from concatenating all the IV-AA-MDS trajectories of the MS subintervals an *MS BV-AA-MDS trajectory*. If an MS BV-AA-MDS trajectory is generated using an initial parameter vector, then the trajectory can be called an *initial MS BV-AA-MDS trajectory*. Two different types of approaches for finding initial parameter vector candidates will be described in subsections of this subsection (subsubsections 3.4.3.5 and 3.4.3.6). And, both of the approaches have been incorporated into the overall strategy outlined in subsection 3.4.4 and applied in subsection 3.4.5 to study conformational transitions for an $\text{N-acetyl-N'}$-methylalaninamide. Both of these approaches work best with the availability of an IV-AA-MDS trajectory that meet some criteria designed to identify trajectories that approximately solve a BVP. We will call such a trajectory an *approximate BV-AA-MDS solution*. The search for an approximate BV-AA-MDS solution can be performed by a random approach, by a strategic approach, or by approaches that have both random and strategic elements. Finding an approximate BV-AA-MDS solution will be the topic of the first four of nine subsubsections that follow. The strategic approach that we will describe uses an approximate method (e.g. one of the methods described in section 3.3.1.2) to generate a BV-AA-$\text{prx}$ trajectory where $\text{prx}$ is a symbolic label for an arbitrary approximate method. This trajectory
satisfies the boundary conditions of the BVP, but, will not, in general, satisfy the differential equation (i.e. the equations of motion). A comparison of a random and strategic approach for generating an approximate BV-AA-MDS solution will be provided in section 3.4.3.1. The ending time and mesh selection will be determined either in the process of finding an approximate BV-AA-MDS solution or later in the process of finding an initial parameter vector.

### 3.4.3.1 Assumptions about simulations with same initial velocity direction

In this section we attempt to describe an assumption that is used in our strategic approach to find an approximate BV-AA-MDS solution. A snapshot of the velocity trajectory is a $3n$-dimensional vector. Like any vector, it has a magnitude and a direction. Suppose that for the IVP

\begin{align}
M \ a(t) &= f(x(t)) , \quad t_0 \leq t_i \leq t_f \\
(3.36) \\
\dot{x}(t_i) &= v_i, \quad v(t_i) = v_i
\end{align}

a solution exists on $[t_0, t_f]$ and it satisfies

\begin{align}
(3.37) \\
\sum_{j=1}^{n} m_j v_{j} (3j - 1)^{+1} &= 0, \\
(3.39) \\
\sum_{j=1}^{n} m_j v_{j} (3j - 1)^{+2} &= 0, \\
\sum_{j=1}^{n} m_j v_{j} (3j - 1)^{+3} &= 0
\end{align}

The equations of (3.39) insure that the linear momentum is zero. Assuming the IVP (3.36), (3.37) is nonlinear, we cannot make precise statements about the solution to the related IVP (3.36),

\begin{align}
(3.40) \\
x_{a}(t_i) &= x_i, \quad v_{a}(t_i) = \alpha v_i
\end{align}

for a real number $\alpha$. However, the methodology that we will use for generation of initial parameter vector candidates is based on the following probabilistic assumptions about the IVP (3.36), (3.40) given the result (3.38)
For $\alpha > 1$, the conditional likelihood that $x_\alpha(t_{\alpha 0}) \approx x_0$ and/or $x_\alpha(t_{\alpha f}) \approx x_f$ given (3.38) is greater than the unconditional likelihood that $x_\alpha(t_{\alpha 0}) \approx x_0$ and/or $x_\alpha(t_{\alpha f}) \approx x_f$. Relationships between the pair $t_{\alpha 0}$ and $t_{\alpha f}$ and the pair $t_0$ and $t_f$ will be specified below. If we make the additional assumption that $x$ is a somewhat linear trajectory in some coordinate system so that $x(t_0) = x_0$ and $x(t_f) = x_f$ on the first passage of $x$ near $x_0$ and $x_f$ respectively, in some appropriate measure of distance, then we hypothesize that the following inequalities tend to be satisfied:

$t_{\alpha 0} > t_0$ and $t_{\alpha f} < t_f$. For $0 < \alpha < 1$, the entire statement holds except that the two inequalities of the previous sentence are reversed so that $t_{\alpha 0} < t_0$ and $t_{\alpha f} > t_f$. The values $t_{\alpha 0}$ and $t_{\alpha f}$ may not be unique. The statement above may be true, for example for $t_{\alpha 0}$ and $t_{\alpha f}$ in intervals near $t_0$ and $t_f$. Additional assumptions are that the ratio of this conditional likelihood to the unconditional likelihood will be greater for $\alpha$ nearer to 1, and, that for $\alpha$ nearer to 1, $t_{\alpha 0}$ will be nearer to $t_0$ and $t_{\alpha f}$ will be nearer to $t_f$. Finally, the closeness of the approximation $x_\alpha(t_{\alpha 0}) \approx x_0$ and/or $x_\alpha(t_{\alpha f}) \approx x_f$ is not specified here. Presumably, this would be done on a case by case basis.

It is worthwhile to reemphasize that the statement above is not mathematically rigorous or precise and certainly will not always hold. It is included as an attempt to provide some intuition into the forthcoming strategic approach outlined in section for generation of initial parameter vector candidates. An illustration of the application of (3.41) follows in subsubsection 3.4.3.2.

### 3.4.3.2 A strategic approach for finding an approximate BV-AA-MDS solution and estimating the time interval for a conformational transition

A strategic approach for finding an approximate BV-AA-MDS solution is described in this subsubsection. In the search for an approximate BV-AA-MDS solution, one can search directly for $s_0 = [x_0; v_0]$, that is, atomic positions and atomic velocities at $t_0$. There are other possibilities, however, due
to the time-reversal symmetry of the Newtonian equations of motion, which is defined and illustrated in [Hol1999]. With respect to the task of finding approximate BV-AA-MDS solutions, the time-reversibility has some useful implications that we now proceed to describe. First, assume that the IVP (3.1), (3.2) is solved and that at \( x(t_f) = x_f \) and \( v(t_f) = v_f \). Then, the IVP

\[
M a(t) = f(x_b(t)), \quad -t_f < t < -t_0
\]

\[
x_b(t_f) = x_f, \quad v_b(-t_0) = -v_f
\]

satisfies \( x_b(-t) = x(t), \quad v_b(-t) = -v(t) \) for \( t_0 < t < t_f \). Note specifically that \( x_b(-t_0) = x_0, \quad v_b(-t_0) = -v_0 \). With respect to the BVP, (3.1), (3.3), (3.28), this implies that an approximate solution of

\[
M a(t) = f(x_b(t)), \quad -t_f < t < -t_0
\]

\[
[x(-t_f) - ALIGN(m,x(-t_0),x_0) : x(-t_0) - ALIGN(m,x(-t_0),x_0)] = 0
\]

is equivalent to an approximate solution of (3.1), (3.3), (3.28). We will refer to the IVP (3.1), (3.2) as a forward IVP (from \( t_0 \) forward in time to \( t_f \)) for the BVP (3.1), (3.3). And, we will refer to the IVP (3.42), (3.43) as the reverse IVP (from \( t_f \) backward in time to \( t_0 \)) for (3.1), (3.3). Assuming boundary conditions are separable as in (3.28), then, in the case of the forward IVP, boundary conditions at \( t_0 \) are satisfied by appropriate choice of initial conditions. An approximate solution approximately satisfies boundary conditions at \( t_f \). By contrast, in the case of the reverse IVP, boundary conditions at \( t_f \) are satisfied by appropriate choice of initial conditions and an approximate solution approximately satisfies boundary conditions at \( t_0 \). More generally, initial conditions can chosen at any \( t_i \) such that \( t_0 < t_i < t_f \). Then, a forward IVP of the form

\[
M a(t) = f(x_i(t)), \quad t_i < t < t_f
\]

\[
x(t_i) = x_i, \quad v(t_i) = v_i
\]

and a reverse IVP of the form

\[
M a(t) = f(x_b(t)), \quad -t_f < t < -t_i
\]
can be concatenated to give an IV-AA-MDS trajectory on \( t_0 < t < t_f \). We will refer to these two IVP’s collectively as a \textit{bi-directional IVP} for the BVP (3.1), (3.3) with initial node \( t_i \). If the concatenated trajectory approximately satisfies boundary conditions at \( t_0 \) and at \( t_f \), then it is an approximate BV-AA-MDS solution.

A conceptual diagram of approximate BV-AA-MDS solutions generated by these three types of IVP’s is provided in Figure 3.4. In this figure, AA-MDS trajectories are indicated by vectors with initial timepoints are the tail of the vector and the ending timepoint is the head of the vector. The boundary conditions are represented by colored splotches. The boundary condition at the beginning of the interval is represented by \( x_0 \) and the boundary condition at the end of the interval is represented by \( x_f \). The BV-AA-MDS solution is the vector colored in black. Approximate BV-AA-MDS solutions generated by a forward IVP, a reverse IVP, and a bi-directional IVP are indicated by blue, maroon, and green vectors, respectively.

The strategic approach assumes existence of a BV-AA-\textit{prx} trajectory for which the linear momentum is zero. (If the velocity vector at time \( t_i \) is \( v_i \), then the linear momentum is zero if (3.39) is satisfied). With the existence of the BV-AA-\textit{prx} trajectory, a series of initial parameter vector candidates could be generated by solving forward IVP’s, reverse IVP’s, and/or bi-directional for initial conditions at any \( t \) in \([t_0, t_f]\). For the description here, we assume we are interested in solving forward IVP’s. Suppose we wish to generate a series of initial parameter vector candidates for the BVP (3.1), (3.3), (2.28). Let \( x^{\text{prx}} \) be a BV-AA-\textit{prx} position trajectory and let \( v^{\text{prx}} \) be the corresponding BV-AA-\textit{prx} velocity trajectory such that the 1st snapshot is \( x^{\text{prx}}(0,:) = x_0 \) and the last of \( N+1 \) snapshots is \( x^{\text{prx}}(N,:) = \text{ALIGN}(m,x_0,x_f) \). The strategy to generate the candidates is to solve a series of forward IVP’s of the form

\[
(3.50) \quad M \dot{a}(t) = f(x(t)), \quad t_0 < t < t_{\text{max}}
\]

\[
(3.51) \quad x(t_0) = x_0, v(t_0) = a(0)v^{\text{prx}}(1,:)
\]
where \( \alpha(i) \) is the scaling factor so that the total energy of the system for the \( i \)th forward IVP has total energy \( H(i) \) where \( H \) is a vector with components representing pre-determined total energy levels for the forward IVP’s. The ending times for the forward IVP’s are set to \( t_{\text{max}} \) under the presumption that the conformational transition of interest would be complete prior to the time \( t = t_{\text{max}} \). So, \( t_f < t_{\text{max}} \).

3.4.3.3 Example 10: Utility of strategic for finding an approximate BV-AA-MDS solution

In this section, the utility of the strategic approach for finding an approximate BV-AA-MDS solution is illustrated using a particular conformational transition for \( N\)-acetyl-\( N'\)-methylalaninamide. The conformational transition of interest can be described as a transition from the primary \( C_{7\text{eq}} \) local minimum to the primary \( C_{5\beta} \) local minimum (i.e. \( C_{7\text{eq}} \rightarrow C_{5\beta} \), using the notation introduced in section 1.3). So, the BVP is of the form (3.1), (3.3), (3.28) with \( x_0 \) corresponding to a \( C_{7\text{eq}} \) local minimum conformation and \( x_f \) corresponding to local minimum conformation of \( C_{5\beta} \). While we assume \( t_0 = 0 \), the ending time for the BVP has not been determined. Only forward IVP’s are considered. We assume that \( t_f < t_{\text{max}} \) where \( t_{\text{max}} = (799) \Delta t = (799) (0.015725 \text{ AKMA units}) \approx 0.614 \text{ ps} \). Approximate trajectories were generated using a particular BV-AA-prx method, distance matrix interpolation using an elastic network model (DMI-ENM), that will be discussed in detail in Chapter 5. Nine different total energy levels are considered. In \( \text{kcal mol}^{-1} \), \( H = [-21; -19; -17; -15; -13; -8; -3; 10; 20] \).

The results for the nine forward IVP’s are displayed using a series of \( \phi-\psi \) plots in Figure 3.6. Inspection of these plots suggests that, qualitatively, the three IVP’s with the lowest total energy do not pass near the \( C_{5\beta} \) local minimum. However, the six trajectories of highest total energy all move from the \( C_{7\text{eq}} \) local minimum toward the \( C_{5\beta} \) energy wells, thereby passing near the \( C_{5\beta} \) local minimum. For comparison purposes, Figure 3.6 shows a series of \( \phi-\psi \) plots from trajectories generated by forward IVP’s with the beginning conformation of the \( C_{7\text{eq}} \) local minimum and with initial velocities randomly sampled from a normal distribution and scaled so that \( H = -13 \text{ kcal mol}^{-1} \). The duration of each of these nine simulations is \( t_{\text{max}} \) as given above. This series of figures suggests that the transitional event occurs
relatively rarely at this energy level. However, we were able to generate an approximate solution at this energy level and several other energy levels using the strategy described in the previous section.

3.4.3.4 Quantitative criteria for identification of approximate BV-AA-MDS solution

We have appealed to \(\varphi-\psi\) plots to qualitatively analyze trajectories. There are three types of quantitative criteria that we use to identify an approximate BV-AA-MDS solution. In genera, let \(D+1\) represent the number of elements of the mesh of an IVP. In particular, here let the IVP be a forward IVP of the form (3.50),(3.51). Let \(x^{\text{ip}}\) be the \(3n \times (D+1)\) position trajectory that is a solution of the forward IVP.

The 1st type of criteria is a requirement that the RMSD from the desired ending structure is less than some value, \(R_{\text{TOL}}\). The value of \(t_f\) for the BVP has yet to be defined. It could be any value between \(t_0\) and \(t_{\text{max}}\). It may be defined in the process of analyzing the appropriate RMSD values. In particular, the following equations specifies the requirement

\[
(3.52) \quad \text{RMSD}_{\text{min}}(m, x(\Delta^i), x_f) \equiv \min_{t_0 \leq t_i \leq t_{\text{max}}} \text{RMSD}(m, x(t_i), x_f) \leq R_{\text{TOL}},
\]

\[
(3.53) \quad t_f(\text{RMSD}_{\text{min}}) = \arg\min \text{RMSD}_{\text{min}}(m, x(\Delta^i), x_f) = \arg\min_{t_0 \leq t_i \leq t_{\text{max}}} (\text{RMSD}(m, x(t_i), x_f)).
\]

The latter equation stipulates that the ending time is assigned to the time \(t_i\) that minimizes \(\text{RMSD}(m, x(t_i), x_f)\) for \(t_0 \leq t_i \leq t_{\text{max}}\). The other two types of criteria are restrictions on the total energy, \(H\), and additional restrictions on the ending time, \(t_f\):

\[
(3.54) \quad H(x,v,m) \leq H_{\text{MAXTOL}}.
\]

\[
(3.55) \quad t_f \leq t_{\text{MAXTOL}}.
\]

There are, at least, three reasons for placing a restriction on \(H(x,v,m)\). First, a system with high energy would be expected to be in constant state of transition, so that a conformational transition may be a frequent event. There is greater interest in the study of conformation transitions that are relatively rare. The total energy level of a MS BV-AA-MDS solution trajectory tends to be similar to the energy level of the initial MS BV-AA-MDS trajectory. So, an upper bound may be placed on the total energy level of the initial MS BV-AA-MDS trajectory to encourage MS BV-AA-MDS solutions of lower total energy. Secondly, there is
a relationship between total energy and temperature due to a direction relationship between kinetic energy and temperature. So, solutions with high total energy may correspond to solutions at temperatures for which the molecules of interest would be denatured and therefore may not be relevant to the physical process being studied. Lastly, low frequency motions tend to correspond to globally oriented, coordinated changes of large amplitude ([Hin2004], [Tam2001], [Kun2004], [Son2006]). In contrast, higher frequency motions tend to be more local in nature. Low frequency motions also require less energy than high energy motions. So, a conformational transition accomplished at a lower total energy is likely to include more low frequency motions and less high frequency motions. This implies that a lower energy conformational transition is more likely to be limited to motions essential to the transition. Analysis of lower energy transition trajectories could be enlightening. Also, for a fixed total energy, [Bai2005] provide anecdotal evidence to suggest that there is a minimal time for transition. Transitions that take less than this minimal time would not be feasible at that energy level. All transitions that approximately take this minimal time bear similar properties in the authors’ analysis in [Bai2005]. Transitions that take significantly longer than this minimal time may not be unique. It could be argued that the unique minimal time transitions are more likely to include just the essential elements of the transition. Regardless, conformational transitions of different durations may exhibit different properties, so depending on the motivations of the analyst, criteria for the ending time, $t_f$, may be appropriate.

### 3.4.3.5 Finding initial parameter vector candidates by directly using an approximate BV-AA-MDS solution

Once an approximate BV-AA-MDS solution has been generated, the atomic positions and atomic velocities from the approximate BV-AA-MDS solution at each of the mesh time points can be inserted into $s^0$. So, if $x^{op}$ and $v^{op}$ are the $3n \times (D+1)$ position and velocity trajectories, respectively of the approximate BV-AA-MDS solution, respectively, $\Delta = [t_0, t_1', t_1, ..., t_N']$, $x^0$, $v^0$ are the $(N+1) \times I$ mesh of MS shooting points, a $3n \times N$ matrix of initial coordinates at the shooting points, and a $3n \times N$ matrix of initial velocities at the shooting points, then we can write using MATLAB notation.
(3.56) \[ x^e = x^{hp}(\cdot, \Delta^e(I:N)) \text{ and } v^e = v^{hp}(\cdot, \Delta^e(I:N)) \]

An initial parameter vector can be generated by this method for any choice of \( \Delta^e \). Choosing \( \Delta^e \) so that the mesh points are uniformly distributed is one possibility and is one that we have employed in the case studies of this chapter. Another factor may be used in choice of \( \Delta^e \) is the conditioning of the sensitivity matrices, \( Y_i \), of (3.18). The conditioning and condition number of the sensitivity matrices is discussed in subsection 2.4.3. The condition numbers of these sensitivity matrices are the condition numbers for the IVP’s of the MS subintervals. In order for the MS method to be effective, as was argued in subsection 2.4.3, none of these condition numbers should be so large that the matrix is ill-conditioned. Below the ill-conditioning threshold, if the condition number were a good predictor of the usefulness of the MS method, one could consider designing an algorithm for selecting \( \Delta^e \) so that the condition number was uniformly low for the initial MS BV-AA-MDS trajectory. Experiments to test this idea have not yet been conducted by the author, but could be considered for future research.

In this approach, the initial MS BV-AA-MDS trajectory will be an AA-MDS trajectory. Figure 3.7 and Figure 3.8 contain conceptual diagrams to illustrate this approach. These figures also illustrate another approach which will be described in the next subsubsection. A more complete description of the contents of these figures will be provided there.

3.4.3.6 Finding initial parameter vector candidates by directly using a BV-AA-\textit{prx} trajectory

Another approach for finding initial parameter vector candidates is to use an approximate method (e.g. one of the methods described in section 3.3.1.2) to generate a BV-AA-\textit{prx} trajectory on the mesh, \( \Delta^e \), where \textit{prx} is a symbolic label for an arbitrary approximate method. This trajectory satisfies the boundary conditions of the BVP, but, will not, in general, satisfy the differential equation (i.e. the equations of motion). From this trajectory, atomic positions and atomic velocities for each of the mesh time points can be inserted into \( s^0 \). So, if \( x^{\textit{prx}} \) and \( v^{\textit{prx}} \) are the \( 3n \times N \) position and velocity trajectories, respectively of the approximate BV-AA-MDS solution, respectively, \( \Delta^e = [t_0^e; t_1^e; t_2^e; \ldots ; t_N^e]^T \), \( x^0 \) are the \( (N+1) \times 1 \) mesh of
MS shooting points, a \(3n \times N\) matrix of initial coordinates at the shooting points, and a \(3n \times N\) matrix of initial velocities at the shooting points, then we can write using MATLAB notation

\[
(3.57) \quad x^0 = x^{\text{prx}}(:, 1:N) \text{ and } v^0 = v^{\text{prx}}(:, 1:N)
\]

An initial parameter vector can be generated by this method for any choice of \(\Delta'\) as long as a BV-AA-\textit{prx} trajectory can be generated for \(\Delta'\). Note that upon applying \(s^0\) to solve IVP's on the MS subintervals the initial MS BV-AA-MDS trajectory will not, in general, be an AA-MDS trajectory since there may be discontinuities in atomic positions or velocities at the internal mesh points.

In the next section there is an illustration of this idea for an example. Additionally, conceptual diagrams of a BV-AA-MDS solution, a BV-AA-\textit{prx} trajectory, and resulting initial MS BV-AA-MDS trajectories generated by the BV-AA-\textit{prx} method and the approximate BV-AA-MDS solution method are provided in Figure 3.7 and Figure 3.8. These figures differ only by choice of mesh. The choice of mesh can have an effect on the magnitude of the initial residual vector (and potentially the eventual convergence) as is illustrated in these figures. The residual vector is smaller for the choice of mesh for Figure 3.7 than for the choice of mesh for Figure 3.8. In these figures, AA-MDS trajectories are indicated by vectors with initial timepoints are the tail of the vector and the ending timepoint is the head of the vector. The boundary conditions are represented by colored splotches. The boundary condition at the beginning of the interval is represented by \(x_0\) and the boundary condition at the end of the interval is represented by \(x_f\). The BV-AA-MDS solution is the vector colored in black. A BV-AA-\textit{prx} solution is represented by a curved orange line. For the MS method with \(N=3\), IVP’s on the subintervals are represent by blue vectors of varying shades. The magnitude of the residual vector is represented conceptually by the light blue vertical bars. For comparison purposes, an approximate BV-AA-MDS solution generated by a forward IVP is indicated by a dashed blue vector. The magnitude of the residual vector is represented conceptually by the magnitude of the navy blue vertical bar.

In approaches for generating a BV-AA-\textit{prx} trajectory, we can consider each snapshot of the position trajectory to be a model of the snapshot of a BV-AA-MDS trajectory at a point in time between the
initial time, \( t_0 \), and the ending time, \( t_f \). However, in some approaches, the model does not include an estimate of the time, \( t \), for a given snapshot. We will call a trajectory with this property a \textit{hidden time trajectory} and we will call an approach or method with this property a \textit{hidden time approach} or \textit{hidden time method}. Just like the case where the components of \( \Delta' \) are explicitly defined, the initial parameter vector, \( s^0 \), is comprised of \( x^0 \) and \( v^0 \) as defined in (3.57). However, the components of \( \Delta' \) are unknown. In the next section, a strategy is given for selection of the components of \( \Delta' \), including the ending time, \( t_N = t_f \).

Any physical properties that are characteristic of solutions to (3.1) where the force field is determined by a potential energy function of the form (3.35) will be characteristic of BV-AA-MDS solution trajectories generated by a numerical method (within limits of accuracy of numerical solution). It is reasonable to assume that, other things being equal, an initial MS BV-AA-MDS trajectory that exhibits these characteristic properties is preferred over one that doesn’t. The physical properties that are of particular interest are (1) conservation of linear momentum, (2) conservation of angular momentum, (3) susceptibility to global rotation, and (4) conservation of total energy.

We seek initial MS BV-AA-MDS trajectories that exhibit these properties. Not all BV-AA-prx trajectories satisfy these properties. So, to generate an initial MS BV-AA-MDS trajectory that satisfies these properties may require modifications of the BV-AA-prx trajectory. The strategy to be described in the next section for generating an augmented initial parameter set using a hidden time BV-AA-prx trajectory also includes methods for modifying a BV-AA-prx trajectory so that the physical properties described above are accommodated.

3.4.3.7 Finding an augmented initial parameter set using a hidden time BV-AA-prx trajectory

As indicated in the previous subsection, this subsection contains a strategy to be described for generating an augmented initial parameter set using a hidden time BV-AA-prx trajectory that satisfies some additional conditions. The additional conditions are that (1) the center of mass of the system is always \([0;0;0]\) — the origin; (2) the linear momentum is constant and is equal to \([0;0;0]\) — the zero vector; (3) for
each $i$ such that $1 \leq i \leq N-1$, the orientation of the initial position for the IVP on subinterval $[t_i^s, t_{i+1}^s]$ are optimal in terms of a RMSD calculation with the ending position on the subinterval $[t_i^s, t_i^e]$, and (4) the total energy is constant. Furthermore, the resulting augmented initial parameter set is intended to produce an initial MS BV-AA-MDS trajectory that has a initial residual vector of relatively small magnitude. As in previous sections, assume that $s^0$ is composed of $x^0$ and $v^0$ where

$$x^0 = [x_0^0; x_1^0; \ldots; x_{N-1}^0], \quad v^0 = [v_0^0; v_1^0; \ldots; v_{N-1}^0],$$

And, assume that we have already made the assignments

$$x_i = x^{prx}(t_i), \quad v_i = v^{prx}(t_i), \quad 0 \leq i \leq N-1,$$

We will describe possible revisions to this assignment below.

The center of mass of a system at the yet to be determined time $t_i^s$ then is given by

$$x_{i}^{cm} = [x_{i1}^{cm}; x_{i2}^{cm}; x_{i3}^{cm}]$$

where

$$x_{i1}^{cm} = \sum_{j=1}^{n} m_j x_{i(3j-1)+1}$$

$$x_{i2}^{cm} = \sum_{j=1}^{n} m_j x_{i(3j-1)+2}$$

$$x_{i3}^{cm} = \sum_{j=1}^{n} m_j x_{i(3j-1)+3}$$

Since all the components of the relevant force field (3.35) are defined in terms of internal coordinates of the system, the center of mass of the system at $t = t_0$ in rectangular coordinates can be assigned arbitrarily. We choose the center of mass to be the origin. And, since the linear momentum will be set equal to the zero vector, the center of mass of the system will always be the origin, in theory. If the BV-AA-prx trajectory used to create the initial parameter set does not satisfy this property, an initial parameter set which does satisfy this property can be created by performing, where necessary, the replacement

$$x_i \mapsto x_i - \{ x_i^{cm} \}^n$$

where $\{ x_i^{cm} \}^n = [ x_i^{cm}; x_i^{cm}; \ldots; x_i^{cm} ]$ is a $3n \times 1$ vector containing the $3 \times 1$ vector $x_i^{cm}$ repeated $n$ times.

For an isolated system, there are no external forces in the force field, so the linear momentum of a system governed by (3.1) is constant. The linear momentum of the system at time $t_i^s$ is given by
To ensure that the linear momentum will be equal to the zero vector, an initial parameter set which does not satisfy this property can be modified by performing the replacement

\[
\begin{align*}
(3.63) \quad v_i & \mapsto v_i - \{ p_{i^{by}} \}^n \\
\end{align*}
\]

where \( \{ p_{i^{by}} \}^n = [ p_1^{by}, p_2^{by}, \ldots, p_i^{by} ] \) is a \( 3n \times 1 \) vector containing the \( 3 \times 1 \) vector \( p_i^{by} \) repeated \( n \) times.

Having described how the 1\(^{\text{st}}\) two conditions are met, we next describe how the components of \( \Delta^i \) are assigned before describing how 3\(^{\text{rd}}\) and 4\(^{\text{th}}\) conditions are met. The idea behind the assignment of values to the components of \( \Delta^i \) is that the hidden time BV-AA-prx trajectory is a good approximation to a BV-AA-MDS solution trajectory in that the following two assumptions are satisfied. The first assumption is that the sequence of conformation of the BV-AA-prx position trajectory has at least one matching sequence of conformations on some BV-AA-MDS solution position trajectory pathway such that each conformation of the latter sequence is measurably similar to a corresponding conformation of the BV-AA-prx sequence. The second assumption is that for each snapshot of the hidden time BV-AA-prx trajectory, a forward IVP solved with initial conditions from that snapshot will result in an IV-AA-MDS position trajectory that moves toward a conformation that approximates the conformation of the next snapshot of the hidden time BV-AA-prx trajectory. More specifically, consider the sequence of IVP’s

\[
(3.64) \quad \mathbf{M} \mathbf{a}(t) = \mathbf{f}(\mathbf{x}(t)) , \quad 0 < t < t_{\max} \\
(3.65) \quad \mathbf{x}(t_i) = \mathbf{x}^{prx}(t_i) , \quad v(t_i) = \mathbf{v}^{prx}(t_i)
\]

for \( 0 \leq i \leq N \). For each \( i \), the position trajectory begins in the \( x^{prx}(t_i) \) conformation and the initial velocity is \( v^{prx}(t_i) \). We make the assumption that the evolution of the IVP results in a position trajectory that moves
closer, in a measurable sense using a metric like $RMSD$, to the $x^{prx}(t_{i+1})$ trajectory monotonically for period of time until a time $t=\Delta t^{i}$ for which reaches a local minimum that satisfies

\begin{equation}
RMSD(m, x(\Delta t^{i}), x^{prx}(t_{i+1})) \leq RMSD(m, x(t), x^{prx}(t_{i+1})),
\end{equation}

for all $t < t_{\text{max}}$. Then, the components of $\Delta^{i}$ can be determined by the formula $t_{i+1}^{i} = t_{i}^{i} + \Delta t^{i}$ for $0 \leq i \leq N - 1$.

Provided the assumptions described above are satisfied, the above description is sufficient to determine $\Delta^{i}$. With respect to the conditions on the orientation of the molecule, it would seem reasonable to consider using the conservation of angular momentum to determine the appropriate orientation. However, it was reported in [Zho2000] that a flexible molecule like a protein can exhibit global rotation even when the angular momentum is zero (perhaps somewhat like a cat exhibiting instinctive acrobatics when falling so that it can land on its feet). Because the sequence of IVP’s of the form (3.64), (3.65) will have been performed and $\Delta^{i}$ has been determined, however, the state of the system at the end of each subinterval is known. So, the initial position of the molecule at the beginning of $[t_{i}^{i}, t_{i+1}^{i}]$ can be optimally aligned with the orientation of the molecule at the end of $[t_{i-1}^{i}, t_{i}^{i}]$ by performing the replacement

\begin{equation}
\begin{bmatrix}
\alpha \\
\eta
\end{bmatrix}
\end{equation}

So, the BV-AA-$prx$ structure at $t_{i}^{i}$ is rotated so that it is optimally superimposed onto the location on the IV-AA-MDS trajectory for $[t_{i-1}^{i}, t_{i}^{i}]$ at $t = t_{i}^{i}$. To preserve the essentials of the BV-AA-$prx$ trajectory and presumably maintain direction toward the next target structure, $x^{i+1}$. we also rotate $v_{i}$ by the same rotation matrix that was used in the replacement (3.67). Because of the use of $m$ as the weight vector, rotation of the velocities conserves linear momentum. And, we can also scale $v_{i}$ to maintain conservation of energy. Since the linear momentum is zero, the rescaling of $v_{i}$ does not affect linear momentum. So, the velocity at $v_{i}$ is computed as

\begin{equation}
(3.68) v_{i} \mapsto \eta \begin{bmatrix}
\alpha \\
\eta
\end{bmatrix} R_{\text{min}}^{\alpha} \left( m, \bar{x}_{i-1}(t_{i}^{i}) - \bar{x}_{i-1}(t_{i}^{i}), \bar{x}_{i}(t_{i}^{i}) - \bar{x}_{i}(t_{i}^{i}) \right)
\end{equation}

where $\alpha$ is chosen to maintain energy conservation.
3.4.3.8 Example 11: Finding an augmented initial parameter set using a hidden time BV-AA-prx trajectory

The assignment of the values of $\Delta s$ described in the previous section will be illustrated with a specific conformational transition of $N$-acetyl-$N'$-methylalaninamide in vacuo. In terms of the $\phi$-$\psi$ adiabatic potential energy surface described in section 3.4.2, the transition is from the primary local minimum of the $C7_{eq}$ well to the primary local minimum of the $C6$ well. In this example, a BV-AA-prx trajectory exists with six conformations including the beginning and ending conformations. (It was generated by distance matrix interpolation which will be described in Chapter 5.) So, we consider applying MS with five subintervals (i.e. $N=5$). A constant time step velocity Verlet IVP solver with the time step, $\Delta t$, set to $\Delta t = 0.015725$ AKMA units. We seek a conformational transition over a duration of less than $125$ fs. This constraint can be implemented by setting $t_{\text{max}} = (48.88821 \times 10^{-15}$ fs / AKMA units) $\times (2.547612$ AKMA units) $= 124.5782268$ fs $= 162(0.76876710225)$ fs. So the maximum number of steps is 162 and the maximum number of snapshots in the initial MS BV-AA-MDS trajectory is 163.

The sequence of five IVP's of the form (3.64), (3.65), were solved resulting in $\Delta t^i_s = 29 \Delta t$ fs, $\Delta t^i_s = 22 \Delta t$ fs, $\Delta t^i_s = 37 \Delta t$ fs, $\Delta t^i_s = 41 \Delta t$ fs, and $\Delta t^i_s = 32 \Delta t$ fs. So, $\Delta s = [29; 51; 88; 129; 161] \Delta t$ fs. In Figure 3.9, the weighted RMSD from the initial conformation of the next shooting point, $RMSD(m, x(t), x^{\text{prx}}(t))$, is calculated and plotted above for each trajectory. This plot shows that qualitatively the second assumption described in section 3.4.3.7 is satisfied. To investigate how well the first assumption is satisfied requires information about solution trajectories.

3.4.3.9 Restarting

The complex, high-dimensional nature of AA-MDS and its potentially highly oscillatory trajectories can be problematic for use of MS for BV-AA-MDS. Sometimes, the MS algorithm leads to a near-solution, but does not find appropriate final steps to achieve convergence. In these situations, the near-solution trajectories exhibit small deviations from one or both of the desired boundary conditions.
and/or slight discontinuities at the intermediate shooting points. (Note that the possibility that neither boundary condition will be satisfied exists when the MS initial guess is derived from a bidirectional IVP). This problem can sometimes be resolved by the following approach:

- **a.** If neither boundary condition is satisfied, we shift the near-solution trajectory so that the initial boundary condition is satisfied. This typically has a minor effect on the general path of the resulting trajectory.

- **b.** Solve an IVP using the initial position of the initial boundary condition or ending boundary condition (depending on which boundary condition is nearer to being satisfied) and record when this trajectory achieves a minimum RMSD criterion for the ending boundary condition. Use this information to assign $t_f$ or $t_0$ and re-apply the multiple shooting algorithm using the approximate BV-AA-MDS approach.

- **c.** To achieve final convergence, it also can sometimes be helpful to perturb the total energy level of the system. So, the IVP solution step mentioned above can also be repeated for a selection of arbitrary scaling adjustments to the initial velocity.

We will call trajectories generated as described above, **restart MS BV-AA-MDS trajectories**. In Figure 3.10, conceptual diagrams of a nearly converged MS BV-AA-MDS trajectory with 3 subintervals, four restart MS BV-AA-MDS trajectories generated as described above using different rescalings of the initial velocity for a translated forward IVP, and a BV-AA-MDS solution trajectory are shown. In this figure, AA-MDS trajectories are indicated by vectors with initial timepoints as the tail of the vector and the ending timepoint as the head of the vector. The boundary conditions are represented by colored splotches. The boundary condition at the beginning of the interval is represented by $x_0$ and the boundary condition at the end of the interval is represented by $x_f$. The BV-AA-MDS solution is the vector colored in black. For the MS method with $N=3$, IVP’s on the subintervals are represent by blue vectors of varying shades. The restart approximate BV-AA-MDS solutions generated by forward IVP’s are indicated by dashed blue vectors.
3.4.4 A MS BV-AA-MDS strategy for min→min transitions

The framework for the physical problem of interest and the relevant numerical methods, and some important concepts for practical application have been provided. Now we set out to describe a comprehensive strategy for finding an augmented initial parameter set and applying multiple shooting methods to BV-AA-MDS. We again consider the BVP of the form (3.1), (3.3), (3.28). To simplify notation, assume that the center of mass for each of the boundary conformations, x₀ and xₖ, is the origin and that xₖ = ALIGN( m, x₀, xₖ). The strategy is outlined below.

A MS BV-AA-MDS strategy for min→min transitions:

1. **Finding approximate BV-AA-MDS solutions.** For a higher probability of eventual convergence to a BV-AA-MDS solution, it is desirable to begin with an initial parameter vector that is derived from an approximate BV-AA-MDS solution. So, a reasonable 1st step is to attempt to find a set of approximate BV-AA-MDS solutions.

   a. **Identification of approximate BV-AA-MDS solutions.** If the BVP of interest has been previously studied and results have been published, then existing literature can be helpful in contain useful information for establishing criteria to identify an approximate BV-AA-MDS solution. In the absence of previously established criteria, one may develop some rough, possibly initially qualitative, means of ranking IV-AA-MDS trajectories based on their nearness to a BV-AA-MDS solution trajectory. In the process of analyzing the IV-AA-MDS trajectories and developing a ranking system, one can identify some quantitative rules for identifying approximate BV-AA-MDS solutions.

   b. **Strategies for finding approximate BV-AA-MDS solutions.** The important things to consider in generating these IV-AA-MDS trajectories is the length of the time interval for the simulations; the point on the time interval for placement of the initial conditions; the initial position of the system; the direction of the initial velocities; and the total energy level of the system, or equivalently, the magnitude of the initial velocities.
2. **Selection of approximate BV-AA-MDS solutions.** Select a set of approximate BV-AA-MDS solutions that meet established criteria.

3. **Selection of number of MS subintervals.** Determine the number of subintervals for the multiple shooting algorithm (i.e. assign a value to $N$).

4. **Selection of optimization methods and parameters.** Select an optimization method and related parameter settings such as convergence criteria and a maximum number of iterations.

5. **Determination of initial MS BV-AA-MDS trajectory.** For each selected approximate BV-AA-MDS solution, generate an augmented initial parameter set. For example, the augmented initial parameter set may be generated by directly using an approximate BV-AA-MDS solution, by directly using a BV-AA-$prx$ trajectory, or by using a hidden time BV-AA-$prx$ trajectory. If possible, the augmented initial parameter set should conserve linear momentum, conserve total energy, maintain consistent orientation at nodes between subintervals. Furthermore, the resulting augmented initial parameter set is intended to produce an initial MS BV-AA-MDS trajectory that has an initial residual vector of relatively small magnitude.

6. **Apply MS algorithm.** For each unique augmented initial parameter set generated in the previous step, attempt to solve BVP using the multiple shooting algorithm.

7. **Restart and re-apply.** If there are MS BV-AA-MDS that are nearly converged, generate restart MS BV-AA-MDS trajectories and re-apply the MS algorithm as described in the previous step.

### 3.4.5 Applying the MS strategy to conformational transitions of $N$-acetyl-$N'\text{-methylalaninamide}$

In this section, we will consider three particular transitions between local minima of the potential energy surface of $N$-acetyl-$N'$-methylalaninamide in vacuo. BVP’s with beginning and ending conditions corresponding to local minima are labeled as $\text{min to min}$ BVP’s, or $\text{min} \rightarrow \text{min}$ BVP’s. In terms of the $\phi$-$\psi$ adiabatic potential energy map described in section 3.4.2, these transitions can be described as transitions
from the $C_7_{eq}$ primary local minimum to the (a) $C_6$ primary local minimum, (b) $C_5_{\beta}$ primary local minimum, (c) $C_7_{ax}$ primary local minimum. The notation $C_7_{eq} \rightarrow C_6$, $C_7_{eq} \rightarrow C_5_{\beta}$, $C_7_{eq} \rightarrow C_7_{ax}$ for transitions (a), (b), and (c), respectively, can also be used. The study of these particular $\text{min} \Rightarrow \text{min} BVP$’s has some intrinsic appeal, but it also is an appealing study as a model for the application of multiple shooting methods to larger peptides or proteins. These BVP’s also are useful for the study of the convergence behavior of our multiple shooting methods.

Since this is an in vacuo study, interaction between the dipeptide and water molecules are not explicitly included. The presence of water molecules may serve to enhance the stability of local minima of the potential energy surface. Even in a vacuum, though, under appropriate conditions, some local minima of the potential energy surface can be fairly stable. Specifically, for sufficiently low levels of energy in the system, randomly-generated trajectories will only rarely pass from potential energy well to another potential energy well. And, at any level of total energy, randomly-generated trajectories will rarely go from precisely from one local minimum to another. So, for the problems we are studying in this chapter, a BV-AA-MDS solution trajectory is a rare trajectory. In an attempt to relate these BVP studies to the actual transitions, we conjecture that solutions to a BVP of this type might identify low energy pathways between starting and ending conformations that are frequently populated by trajectories that transition from one potential well to another. If desired, perhaps one could use a BV-AA-MDS solution trajectory generated by the studies here to generate an ensemble of well to well trajectories using the methods described in “Reaction coordinates of biomolecular isomerization” by Bolhuis, Dellago, and Chandler ([Bol2000]).

Because there is a preference for BV-AA-MDS solution trajectories with lower total energy, we have experimented with BVP’s that have specific conditions that require solution trajectories that have total energy is equal to or below a specified value. The preference for BV-AA-MDS solution trajectories with lower total energy can be accommodated without these additional conditions by selecting initial MS BV-AA-MDS trajectories with lower total energy. We have experienced greater success with the latter
strategy. Following is a summary of the results for the three transitions, organized based on the general strategies outlined in the previous section.

1. **Finding approximate BV-AA-MDS solutions.** BV-AA-prx trajectories were generated by all-atom distance matrix interpolation using an elastic network model (AA-DMI-ENM). This method is described in Chapter 5. To begin, for each of the three transitions, only forward IVP’s are considered. For each transition, we consider nine different total energy levels. The total energy in $kcal mol^{-1}$ levels are the components of the vector $\mathbf{H}$ where

$$\mathbf{H} = [-21; -19; -17; -15; -13; -8; -3; 10; 20]$$

A velocity Verlet IVP solver with the constant time step $\Delta t = 0.015725$ AKMA units was used. After analyzing some preliminary exploratory simulations, the maximum time, $t_{\text{max}}$, was set to $t_{\text{max}} = (799) \Delta t = (0.015725) \text{AKMA units} \approx 0.614 \text{ps}$. For initial conditions derived from AA-DMI-ENM trajectories for $C_{7\text{eq}} \rightarrow C6$ and $C_{7\text{eq}} \rightarrow C5_{\beta}$, the forward IVP’s result in trajectories which appear to move from the $C_{7\text{eq}}$ local minimum to the C6 and C5$_{\beta}$ wells, respectively. For initial conditions derived from AA-DMI-ENM trajectories for $C_{7\text{eq}} \rightarrow C7_{\text{ax}}$, the forward IVP’s do not result in trajectories which move to the $C7_{\text{ax}}$ well. As a result, we also consider a sequence of bidirectional and reverse IVP’s for $C_{7\text{eq}} \rightarrow C7_{\text{ax}}$ with initial conditions based on local data at 25 snapshots along the aforementioned BV-AA-prx trajectory. For each of these we consider four different energy levels, $-8$, $-3$, $10$, and $20$ kcal mol$^{-1}$. Several of these IVP’s did result in trajectories which moved from the $C7_{\text{eq}}$ energy minimum to the $C7_{\text{ax}}$ well.

2. **Selection of approximate BV-AA-MDS solutions.** Let the AA-MDS trajectories be represented by $x^{\text{MDS}}(t)$. The value of $\text{RMSD}(m, x^{\text{MDS}}(t), x_j)$ and the corresponding ending times for the initial set of forward IVP’s are shown in Table 3.2(a)-(c). The values of $\text{RMSD}(m, x^{\text{MDS}}(t), x_0)$ and $\text{RMSD}(m, x^{\text{MDS}}(t), x_j)$ and the corresponding simulation times for the selected bidirectional and reverse IVP’s are shown in Table 3.3. The approximate BV-AA-MDS solution trajectories that were selected
are highlighted and labeled in Table 3.2. The selection criteria for the three different transitions are shown below:

(a) \((\text{RMSD}_{\text{min}}(m, x_{\text{MDS}}(\Delta t) x_f) < \text{0.07})\) or \((\text{RMSD}_{\text{min}}(m, x_{\text{MDS}}(\Delta t) x_f) < \text{0.10} \text{ and } H < -10))\) and \(t_f(\text{RMSD}_{\text{min}}) < (400)0.015725\) AKMA units

(b) \((\text{RMSD}_{\text{min}}(m, x_{\text{MDS}}(\Delta t) x_f) < \text{0.20})\) or \((\text{RMSD}_{\text{min}}(m, x_{\text{MDS}}(\Delta t) x_f) < \text{0.30} \text{ and } H > +10))\) and \(t_f(\text{RMSD}_{\text{min}}) < (400)0.015725\) AKMA units

(c) \((\text{RMSD}_{\text{min}}(m, x_{\text{MDS}}(\Delta t) x_0) + \text{RMSD}_{\text{min}}(m, x_{\text{MDS}}(\Delta t) x_f) < \text{0.52})\) or 
\((\text{RMSD}_{\text{min}}(m, x_{\text{MDS}}(\Delta t) x_0) + \text{RMSD}_{\text{min}}(m, x_{\text{MDS}}(\Delta t) x_f) < \text{0.78} \text{ and } H > +10))\) and 
\(t_f(\text{RMSD}_{\text{min}}) < (400)0.015725\) AKMA units

3. Selection of number of MS subintervals. The MS algorithm is applied with \(N = 1, 2, 3, 4, 5, 6\). This leads to \(6 \times 3 = 18\) different applications of the MS algorithm for \(C_{7\text{eq}} \rightarrow C_{6}\) and \(6 \times 4 = 24\) different applications for \(C_{7\text{eq}} \rightarrow C_{5\beta}\) and \(C_{7\text{ax}} \rightarrow C_{7\text{eq}}\).

4. Selection of optimization methods and parameters. The dogleg trust region global convergence scheme was selected were the maximum number of iterations set to 26. For MS runs which required all 26 iterations, we define weak convergence based on the criterion that \(\| M^{1/2} F \|_2 / \sqrt{N} \leq 0.25\) on the final iteration. We define strong convergence based on the criterion that \(\| M^{1/2} F \|_2 / \sqrt{N} \leq 10^{-6}\).

5. Determination of initial MS BV-AA-MDS trajectory. An augmented initial parameter set was generated using a hidden time BV-AA-prx trajectory. The hidden time BV-AA-prx trajectory that was used is an all-atom distance matrix interpolation method using distance geometry with local constrained energy minimization. This method is described in Chapter 5. The method conserves linear momentum, total energy, and maintains consistent orientation at nodes between subintervals. The augmented initial parameter set is intended to produce an initial MS BV-AA-MDS trajectory that has a initial residual vector of relatively small magnitude.
6. **Apply MS algorithm.** Convergence data for each of the $18 + 18 + 24 = 60$ MS applications were collected and analyzed. In many cases, convergence is observed in this step. When strong convergence is achieved, it is usually achieved in 5-12 iterations. Summary data will be described in the next step.

7. **Restart and re-apply.** For each of the three transitions, ten nearly converged trajectories were selected for restarting. For the $20 \text{C}_7 \text{eq} \rightarrow \text{C}_6$ and $\text{C}_7 \text{eq} \rightarrow \text{C}_7 \beta$ transitions, the initial position was chosen to satisfy the $\text{C}_7 \text{eq}$ boundary structure and initial velocities were based on the current iteration’s velocities at $t = t_0$. For the $10 \text{C}_7 \text{eq} \rightarrow \text{C}_7 \text{ax}$ transitions, initial position was chosen to satisfy the $\text{C}_7 \text{ax}$ boundary structure and initial velocities are based on the current iteration’s velocities at $t = t_f$. IV-AA-MDS is performed for the total energy level of the selected trajectory as well as the following energy levels in kcal mol$^{-1}$:

\[
\begin{align*}
\text{C}_7 \text{eq} \rightarrow \text{C}_6: & \{-15, -10, -5, 0, 5, 10, 15, 20, 25\} \\
\text{C}_7 \text{eq} \rightarrow \text{C}_5 \beta: & \{-15, -10, -5, 0, 5, 10, 15, 20, 25\} \\
\text{C}_7 \text{eq} \rightarrow \text{C}_7 \text{ax}: & \{-10, -5, 0, 5, 10, 15, 20, 30, 40\}
\end{align*}
\]

For the twenty $\text{C}_7 \text{eq} \rightarrow \text{C}_6$ and $\text{C}_7 \text{eq} \rightarrow \text{C}_5 \beta$ transitions, $RMSD(m, x^{MDS}(t), x_i)$ and the corresponding values for $t_f$ were recorded. For the ten $\text{C}_7 \text{eq} \rightarrow \text{C}_7 \text{ax}$ transitions, $RMSD(m, x^{MDS}(t), x_i)$ and the corresponding values for $t_0$ were recorded. Three of the ten were selected for each transition. For each of these three, the MS algorithm was applied with $N=1, 2, 3$, and 6 subintervals for a total of $10 \times 3 \times 4 = 120$ applications. Frequencies of convergence, energy levels upon convergence, and transition times are summarized in Table 3.4. An example of convergence data exhibiting strong convergence (with a rate that appears to be super-linear or quadratic) is shown in Table 3.5. For the ensemble of solution trajectories for each transition, the distribution of total energy is shown in Figure 3.11 and $\phi-\psi$ plots are shown in Figure 3.12. Also, as an illustration, $\phi-\psi$ plots for an initial trajectory and a corresponding ending solution trajectory are shown for each type of transition in Figure 3.13.
3.4.6 Additional analysis of \textit{N-acetyl-N'}-methylalaninamide study and further discussion

3.4.6.1 Comparing solution trajectories

The BV-AA-MDS solution trajectories that we have found for the three transitions differ by duration of time interval, or transition time, and total energy level. For a specific solution, the transitions times were always approximately the same as the duration of the approximate BV-AA-MDS solution (shown in Table 3.2 and Table 3.3) that was used for the initial MS BV-AA-MDS trajectory. In spite of these variations, the overlay of the $\phi$-$\psi$ plots for the BV-AA-MDS solution trajectories in Figure 3.12 suggests that all the trajectories followed approximately the same path. Lower energy solution trajectories for $C7_{eq} \rightarrow C6$ and $C7_{eq} \rightarrow C7_{ax}$ followed approximately straight lines on the $\phi$-$\psi$ plots, while higher energy solutions tend to include some deviations from a straight line (Data not shown since in Figure 3.12 trajectories are not distinguishable by energy level). For the $C7_{eq} \rightarrow C5_{p}$ solution trajectories, however, even the lower energy solution exhibit curved lines on the $\phi$-$\psi$ plots. There is some indication of two different types of curves, suggesting two slightly different pathways. Moreover, if other attributes of a trajectory were also considered, it is possible that all of these trajectories could be further categorized into different pathways. There is not a strong correlation between transition time and total energy in our data. However, the shortest durations for $C7_{eq} \rightarrow C6$ did tend to be associated the highest total energy upon convergence. Perhaps for $C7_{eq} \rightarrow C6$, a higher total energy was necessary to make the transition in a short time.

3.4.6.2 Comparing initial trajectories and solution trajectories

In Figure 3.13, for each of the three transitions, a $\phi$-$\psi$ plot is shown for one of the MS BV-AA-MDS solution trajectories and its corresponding initial MS BV-AA-MDS trajectory. Clearly, the evolution of $\phi$ and $\psi$ values for the initial MS BV-AA-MDS trajectory differ from the evolution for the BV-AA-MDS solution trajectory. It is interesting to note that the $\phi$-$\psi$ plot for the initial BV-AA-MDS trajectories are approximately continuous. This might be attributed in part, at least, to the strategy for
determining an initial MS BV-AA-MDS trajectory using the methods described in subsubsection 3.4.3.7. It can also be seen that on some subintervals, the initial trajectory does not seem desirable in the sense that the direction of the $\phi$-$\psi$ evolution for some subintervals is not toward the ending point. These trajectories are evidence that the algorithm may converge even if the initial MS BV-AA-MDS trajectory does not seem ideal.

3.4.6.3 Previously reported analyses of transition between C7$_{eq}$ and C7$_{ax}$ conformations

In [Hu2003] as well as in other sources, it is documented that transitions between C7$_{eq}$ and C7$_{ax}$ are relatively rare and that there appears to be a relatively high energy barrier between C7$_{eq}$ and C7$_{ax}$. The distribution of total energy levels and the minimum total energy level are both considerably higher for the transition between C7$_{eq}$ and C7$_{ax}$ than for the transitions between C7$_{eq}$ and C6 and between C7$_{eq}$ and C5$_{p}$. These differences are consistent with the observations of [Hu2003].

3.4.6.4 Comparing efficiency of BV-AA-MDS and IV-AA-MDS

In Table 3.4, percentages are provided for the number of converged trajectories after the final step was applied in the comprehensive strategy of 3.4.4 was applied. These percentages some idea of the success rate of the strategies we have employed. It should be noted, though, that trajectories were discarded earlier in the process (in step 2 and after step 6). The goal of the MS method to find some representative solutions rather than necessarily find a large ensemble of trajectories. What seems most important is that some solution trajectories can be found. If the efficiency were to be gauged, a relevant statistic seems to be the number of solution trajectories per unit of computing resources.

We consider a comparison of our algorithm with use of an IVP approach in which we repeatedly perform simulations until the desired trajectories occur by chance. Steps 1 and 2 of the general strategies that we have outlined involve generation of approximate BV-AA-MDS solution trajectories in order to generate good initial guesses for the multiple shooting algorithms. We note that these steps would be useful even if we were to use an IVP approach to attempt to solve the BVP. So, it seems appropriate to
consider comparisons of computational costs and efficiencies in steps 3-7 with an IV-AA-MDS. A framework for this type of comparison was developed in section 2.4.4 of this dissertation. Even though one iteration of the MS algorithm in step 6 is more expensive than one IVP solve, overall there may be greater efficiency in terms of the number of solution trajectories per unit of computing resources, if the average number of iterations required for convergence is minimal.

Waiting times for transitions between wells of the potential energy surface of the alanine dipeptide in solution can be on the order of hundreds of picoseconds or nanoseconds ([Che2004]). Even with a good initial guess, it may be that waiting times for transitions between local minima of a potential energy surface of the alanine dipeptide could actually be much longer than waiting times for transitions between wells. On the other hand, the actual transition time between a local minima or between wells can be very fast. In a vacuum, transition times (i.e. actual time from the beginning of the transition to the end) can be only 0.1-0.3 picoseconds and in solution, similar transition times are reported in [Wou2001]. So, the BVP approach may be more efficient since it may be possible to find a solution trajectory much faster than with random IVP methods. As an aside, it is noted that the BVP approach holds out the possibility of using solutions found for one environment (e.g. vacuum) as starting points to rapidly generate solutions for another environment (e.g. aqueous solution ) ([Gil1992]).

3.4.6.5 Satisfying properties of a solution at MS shooting points

We previously pointed out that in the BVP (3.1), (3.3), (3.28), the absolute locations of a molecule at the endpoints are not specified. So, the linear momentum of the system is not specified. Moreover, neither the total energy nor the angular momentum of the system are specified. When the MS algorithm that we have described is applied to this BVP, these quantities are not specified at the MS shooting points either. For a BV-AA-MDS solution trajectory, total energy, linear momentum, and angular momentum must be conserved. For an initial MS BV-AA-MDS trajectory, we have specified in our strategy that the 1st two of these items must be conserved. For an initial trajectory MS BV-AA-MDS trajectory, the angular momentum is not specified, but the ending conformation of one subinterval is required to be optimally
aligned with the beginning conformation of the next subinterval. On the other hand, an intermediate (i.e. not the initial and not a solution) MS BV-AA-MDS trajectory, will, in general, not have any of these properties.

Methods to specifically require or, at least, encourage that these properties are satisfied on each iteration have been considered. With respect to conservation of total energy and conservation of linear momentum, \( N-1 \) additional constraint equations could be added as components of the residual function \( F \).

Some experimental work has not shown this to be an attractive strategy, but, nevertheless, it could be considered more closely in the future. With respect to optimal alignment at each shooting point, this could be incorporated into the algorithm by a modification of \( F \). The Jacobian of \( F \) would change. The block components of \( F' \) that assume the a negative identity matrix in the current implementation would become dense, but would probably be well-approximated by negative identity matrices, particularly near a solution. So, this option could be considered with either the modified Jacobian or with the Jacobian of the unmodified residual vector. The philosophy behind the current implementation is that a initial MS BV-AA-MDS that does satisfy these properties and approximately satisfies the BVP will make convergence to a solution likely. Adding the additional complexities may not necessarily help performance, and could actually be detrimental.

### 3.4.6.6 Reduced number of boundary conditions representing a potential energy well

In Chapter 4, we propose a method for constructing BVP’s with boundary conditions that correspond to entire potential energy wells rather than just one point (e.g. the local minimum of the well). The method is applied for transitions between the \( C7_{\text{eq}} \) and \( C7_{\text{ax}} \) energy wells for the alanine dipeptide. This is a more realistic approach particularly for larger molecules, as we are often more interested in an ensemble of trajectories that show a transition between potential energy wells instead of trajectories that move precisely between local minima. In this method, for the application to transitions between \( C7_{\text{eq}} \) and \( C7_{\text{ax}} \) potential energy wells for alanine dipeptide, a reduced number of boundary conditions is defined based on the requirement that a set of ranges for interatomic distances between non-adjacent heavy atoms are
satisfied for both starting and ending structures. The set of ranges is derived empirically by performing low energy simulations in each potential energy well and observing the ranges of interatomic distances.

### 3.4.6.7 Reduced number of parameters

For application to larger systems, it will also necessary to design an MS algorithm with a computational cost less than that of the algorithm we have described. One possible way may be through a reduction in number of parameters. In Chapter 4, we consider the single shooting case, i.e. $N=1$, with fixed time interval $(0, t_f)$, and a fixed initial conformation which satisfies initial boundary conditions. In this case, the initial position can be fixed, so the $3n$ initial velocities are the only parameters to be determined. A normal mode based approach for finding solutions to an AA-MDS BVP with a reduced number of initial velocities is considered. The directions for this reduced set of initial velocities are the directions of a subset of low-frequency vibrational modes.

### 3.4.6.8 Convergence rate

One of the advantages of initially studying transitions between local minima rather than transitions between potential energy wells is that the BVP is well-posed and convergence properties can be inspected to see if they meet expectations. We expect that on iterations of the global convergence methods employed here for which the parameter set is sufficiently close to a solution, the parameter update for the next iteration will be determined by the Newton step which theoretically will give a quadratic rate of convergence. Due to unidentified factors possibly related to the size of the MS linear system and errors in the numerical solution of ODE’s, quadratic convergence may not always be realized. To achieve strong convergence, i.e. $\| M^{1/2} F \|_2 / \sqrt{N} \leq 10^{-6}$ seems require a rate of convergence approaching quadratic, or, at least, superlinear. An example of rate of convergence is given in Table 3.5.
3.4.6.9 Application of MS to simulation in other ensembles, at lower resolution, and in solution

There is a classification scheme for molecular simulation methods in which a class, usually called an ensemble, is determined by the combination of thermodynamic properties which are theoretically constant during the simulation. In AA-MDS, as it has been defined in this dissertation, an isolated system is subjected to the Newtonian equations of motion. A snapshot of a trajectory is sometimes called a microstate of the system. The trajectory is can be considered to be a sample from the microcanonical ensemble of microstates. In the microcanonical ensemble, the number of particles, the volume, and the total energy are fixed. Simulations are often performed in other ensembles. It would be worthwhile to consider application of MS in other ensembles, such as, for example, the canonical ensemble. This is a popular ensemble in which the temperature is held constant instead of the total energy. There are different methods for simulation in the canonical ensemble.

It may also be useful to consider application of MS to simulation methods that incorporate coarse-grained, or reduced, models of proteins. In [Mal2005], interesting results and physically meaningful results are obtained from molecular dynamics simulations of prion-like proteins performed using a medium resolution lattice model with an empirical potential based on PDB structures. In these reduced models, interactions with water are only included in a highly limited implicit way.

In future research using BV-AA-MDS, we also could consider modeling the interactions between the molecule and the surrounding solvent since this interaction is thought to be critical for transitions or folding processes. A simple way to attempt to model interactions with solvent is to adjust the dielectric constant, $\varepsilon$, in the potential energy function. This adjustment is not thought to be adequate for simulations of protein folding or significant changes in conformation. Still, setting this constant to an appropriate value may be an adequate for simulations of local transitions of an already folded protein. Assessments about appropriate modeling of solvent interactions generally need to be made on a case by case basis ([Cas2002]). Another approach is to simply include explicit solvent molecules in the simulation. Since one must add...
many molecules with this approach, the size of the system can increase significantly. To avoid this
increase in system size, methods have been developed to attempt to model interactions with solvent without
explicitly including solvent. One approach to simulating interaction with water is to use a stochastic
approach and apply Brownian or Langevin dynamics ([Lea2001]). One popular and apparently effective
class of implicit solvent model is the Generalized Born/ Solvent Accessible Surface Area (GB/SA) model.
This method was introduced in [Sti1990] and fast analytical methods for applying this model to molecular
dynamics calculations have been developed ([Qiu1997],[Gho1998]). A GB/SA model was used for
example in the atomistic simulation of the folding of a 16-residue β-hairpin ([Zag2001]).

3.5 Summary

Initial value and boundary value approaches to all-atom molecular dynamics simulation have been
described and the utility as well as limitations of these approaches have been highlighted. The biological
interest in transitions between two known molecular conformations suggest an important application for
boundary value approaches to all-atom molecular dynamics simulation which are designed to
algorithmically seek trajectories that accomplish the transition of interest. The multiple shooting method,
which was previously proposed as a numerical method for this study (in Chapter 2), has been applied in this
chapter to a model problem — transitions between local minima of potential energy wells of
\textit{N-acetyl-N'-methylalaninamide} in \textit{vacuo}. The local minima are specified by sets of internal coordinates.

Important aspects of this endeavor have been described, including strategies for defining boundary
conditions, generating augmented initial parameter sets and possibly modifying them by methods external
to the multiple shooting method, and evaluating convergence. A description of how optimal rotation of
three-dimensional structures can be used in defining boundary conditions and in generating augmented
initial parameter sets has been provided. Some important assumptions that we have used in generating
augmented initial parameter sets have been noted. Ways to use approximate methods for generating
trajectories corresponding to conformational transitions to generate augmented initial parameter sets have
identified. A comprehensive strategy for the use of multiple shooting methods for boundary value
approaches to all-atom molecular dynamics simulation have been presented. As an ancillary comment, some terminology has been introduced with the hope that it contributes simplicity and clarity to the presentation. Perhaps this terminology may evolve to become useful in further related work.

It is shown that application of the strategy for multiple shooting results in generation of solution trajectories, that is, trajectories that accomplish the transition of interest. It would be rare to find a trajectory between local minima by random methods, so solution trajectories are, in some sense, rare by nature. The fact that some of these solution trajectories exhibit low total energy suggest further that trajectories that are qualitatively desirable can be obtained. A quadratic rate of convergence for the iterative multiple shooting method has been observed, suggesting some consistency with theoretical expectations. Solution trajectories and initial trajectories are analyzed with respect to total energy level, duration of time, changes in key dihedral angles of the alanine dipeptide. With respect to direction for future research, more can be learned by further study of the alanine dipeptide and other small molecules. But, the emphasis of choice here is that the alanine dipeptide is a model for larger peptides or proteins. There will be greater complexities and challenges associated with application to larger peptides. It is anticipated that the definition of boundary conditions will need to be modified for larger systems, both to model the physical problem of interest more accurately and to keep the number of boundary conditions manageable. Also, an approach in which the number of parameters of the model is significantly less than six times the number of particles will be necessary for large systems. Before moving to larger systems, methods for accomplishing these objectives are developed and applied to the same alanine dipeptide. This is the primary topic of Chapter 4. Other complexities and challenges that have not been anticipated will undoubtedly arise when applied to larger molecules, but it is expected to be an educational, enjoyable, and worthy pursuit.

3.6 References


Figures and Tables

Table 3.1  Summary data: local minima of N-acetyl-N′-methylalaninamide in vacuo

<table>
<thead>
<tr>
<th>id</th>
<th>conformation</th>
<th>φ</th>
<th>ψ</th>
<th>U</th>
<th>frequency</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>C7\text{eq}</td>
<td>−72</td>
<td>40</td>
<td>−22.7</td>
<td>341</td>
<td>25%</td>
</tr>
<tr>
<td>2</td>
<td>C5\text{β} I</td>
<td>−152</td>
<td>166</td>
<td>−21.8</td>
<td>287</td>
<td>21%</td>
</tr>
<tr>
<td>3</td>
<td>C7\text{ax}</td>
<td>59</td>
<td>−31</td>
<td>−20.4</td>
<td>242</td>
<td>18%</td>
</tr>
<tr>
<td>4</td>
<td>C6</td>
<td>−133</td>
<td>23</td>
<td>−20.1</td>
<td>224</td>
<td>16%</td>
</tr>
<tr>
<td>5</td>
<td>C5\text{β} II</td>
<td>−150</td>
<td>157</td>
<td>−18.6</td>
<td>52</td>
<td>4%</td>
</tr>
<tr>
<td>6</td>
<td>α-R</td>
<td>−72</td>
<td>−13</td>
<td>−17.6</td>
<td>1</td>
<td>0%</td>
</tr>
<tr>
<td>7</td>
<td>α-L</td>
<td>56</td>
<td>20</td>
<td>−13.6</td>
<td>104</td>
<td>8%</td>
</tr>
<tr>
<td></td>
<td>all others</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>63</td>
<td>5%</td>
</tr>
</tbody>
</table>

\[\text{§ Potential energy; \hfill \¶ Number of times found by energy minimization}\]

Table 3.2  Initial trajectory generation, forward direction

<table>
<thead>
<tr>
<th>(a)</th>
<th>(b)</th>
<th>(c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C7\text{eq} \rightarrow C6</td>
<td>C7\text{eq} \rightarrow C5\text{β}</td>
<td>C7\text{eq} \rightarrow C7\text{ax}</td>
</tr>
<tr>
<td>total</td>
<td>RMSD</td>
<td>steps for min</td>
</tr>
<tr>
<td>energy</td>
<td>index</td>
<td>index</td>
</tr>
<tr>
<td>-21</td>
<td>0.51</td>
<td>131</td>
</tr>
<tr>
<td>-19</td>
<td>0.35</td>
<td>217</td>
</tr>
<tr>
<td>-17</td>
<td>0.04</td>
<td>737</td>
</tr>
<tr>
<td>-15</td>
<td>0.10</td>
<td>196</td>
</tr>
<tr>
<td>-13</td>
<td>0.08</td>
<td>163</td>
</tr>
<tr>
<td>-8</td>
<td>0.07</td>
<td>124</td>
</tr>
<tr>
<td>-3</td>
<td>0.06</td>
<td>104</td>
</tr>
<tr>
<td>10</td>
<td>0.06</td>
<td>79</td>
</tr>
<tr>
<td>20</td>
<td>0.07</td>
<td>69</td>
</tr>
</tbody>
</table>

Using initial velocity directions assigned for the C7\text{eq} conformation derived from a distance matrix interpolation trajectory that transitions between C7\text{eq} conformation and C7\text{ax} conformation; zero linear momentum; highlighted rows indicate trajectories selected for generating initial parameter sets for multiple shooting.
Table 3.3  Initial trajectory generation, bidirectional, C7\textsubscript{eq} \rightarrow C7\textsubscript{ax}

<table>
<thead>
<tr>
<th>total energy</th>
<th>initial velocity / conformation index</th>
<th>min RMSD to C7\textsubscript{eq}</th>
<th>min RMSD to C7\textsubscript{ax}</th>
<th>mean RMSD to C7\textsubscript{eq} and C7\textsubscript{ax}</th>
<th>number of forward steps</th>
<th>number of reverse steps</th>
<th>total steps index</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>389</td>
<td>0.29</td>
<td>0.19</td>
<td>0.24</td>
<td>295</td>
<td>225</td>
<td>519</td>
</tr>
<tr>
<td>10</td>
<td>414</td>
<td>0.34</td>
<td>0.15</td>
<td>0.24</td>
<td>259</td>
<td>190</td>
<td>448</td>
</tr>
<tr>
<td>10</td>
<td>410</td>
<td>0.33</td>
<td>0.16</td>
<td>0.25</td>
<td>226</td>
<td>239</td>
<td>464</td>
</tr>
<tr>
<td>20</td>
<td>410</td>
<td>0.33</td>
<td>0.17</td>
<td>0.25</td>
<td>260</td>
<td>198</td>
<td>457</td>
</tr>
</tbody>
</table>

Using initial velocity directions assigned for various intermediate conformation derived from a distance matrix interpolation trajectory that transitions between C7\textsubscript{eq} conformation and C7\textsubscript{ax} conformation; zero linear momentum; highlighted rows indicate trajectories selected for generating initial parameter sets; data only shown for trajectories used later as initial guess trajectories; zero linear momentum.

Table 3.4  Summary data: convergence for min\rightarrow min BVP's

<table>
<thead>
<tr>
<th>transition</th>
<th>$|M^{1/2}F|_2\sqrt{N} \leq 0.25$ §</th>
<th>$|M^{1/2}F|_2\sqrt{N} \leq 1e^{-6}$ ¶</th>
<th>transition time</th>
<th>min total energy</th>
</tr>
</thead>
<tbody>
<tr>
<td>C7\textsubscript{eq} to C6</td>
<td>118/120=98.3%</td>
<td>55/120=45.8%</td>
<td>0.10 - 0.23 ps</td>
<td>-13 kcal</td>
</tr>
<tr>
<td>C7\textsubscript{eq} to C5\beta</td>
<td>88/120=73.3%</td>
<td>10/120=8.3%</td>
<td>0.10 - 0.30 ps</td>
<td>-6 kcal</td>
</tr>
<tr>
<td>C7\textsubscript{eq} to C7\textsubscript{ax}</td>
<td>74/120=61.7%</td>
<td>9/120=7.5%</td>
<td>0.10 - 0.30 ps</td>
<td>7 kcal</td>
</tr>
</tbody>
</table>

§percentage of trajectories that converged at least weakly (in step 7).
¶percentage of trajectories that converged strongly (in step 7).
### Table 3.5  Example of convergence data for $\min \rightarrow \min$ BVP's

<table>
<thead>
<tr>
<th>Algorithm: dogleg trust region</th>
<th>initial energy level (kcal)</th>
<th>$-13$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$, number of steps</td>
<td></td>
<td>163</td>
</tr>
<tr>
<td>$N$, number of subintervals</td>
<td></td>
<td>5</td>
</tr>
</tbody>
</table>

| iteration | $||M^{1/2}F||_2$ | $H$     |
|-----------|-----------------|---------|
| 1         | 2.7868764       | -13     |
| 2         | 0.8215599       | -10.274 |
| 3         | 0.1194558       | -12.49545819 |
| 4         | 0.0143043       | -12.5825488 |
| 5         | 3.45E-06        | -12.58248875 |
| 6         | 1.95E-11        | -12.58248665 |
Figure 3.1  Ball-and-stick visualization of \( N\text{-acetyl-} N'\text{-methylalaninamide} \)

Ball-and-stick visualization of \( N\text{-acetyl-} N'\text{-methylalaninamide} \) Atoms are shown as balls; covalent bonds are shown as sticks. The colors of the atoms represent the type of atom. (hydrogen—white; carbon—greenish-gray; nitrogen—blue; oxygen—red).
Figure 3.2  φ-ψ contour plot: local minima of N-acetyl-N'-methylalaninamide

Marked in the φ-ψ contour plot are six primary local minima of the alanine dipeptide potential energy surface.
Figure 3.3  Ball-and-stick visualization: local minima of N-acetyl-N'-methylalaninamide

Ball-and-stick visualization: local minima of N-acetyl-N'-methylalaninamide  Shown in the picture are the conformations corresponding to the four most frequently observed ending conformations in an empirical approach for identifying minima of the potential energy surface. Clockwise from the top left, the conformations shown correspond to the C6, C5β, C7ax, and C7eq conformations.
Figure 3.4  BV-AA-MDS solution and three approximate BV-AA-MDS solutions

Conceptual diagram of a BV-AA-MDS solution and three approximate BV-AA-MDS. AA-MDS trajectories are indicated by vectors with initial timepoints are the tail of the vector and the ending timepoint is the head of the vector. The boundary conditions are represented by colored splotches. The boundary condition at the beginning of the interval is represented by $x_0$ and the boundary condition at the end of the interval is represented by $x_f$. The BV-AA-MDS solution is the vector colored in black. Approximate BV-AA-MDS solutions generated by a forward IVP, a reverse IVP, and a bi-directional IVP are indicated by blue, maroon, and green vectors, respectively.
Figure 3.5  **Strategic initial parameter vector selection for $C_{7_{eq}} \rightarrow C_{5_{\beta}}$**
Forward IVP’s; BV-AA-DMI-ENM initial velocity direction
Simulation time : 0.614 ps per IVP
Figure 3.6  Forward IVP’s from C7\textsubscript{eq} local minimum; random initial velocity selection
Initial velocities sampled from a normal distribution (μ=0)
\( H. = -13 \text{ kcal mol}^{-1} \); simulation time : 0.614 ps per IVP
Conceptual diagram of a BV-AA-MDS solution and three approximate BV-AA-MDS. AA-MDS trajectories are indicated by vectors with initial timepoints are the tail of the vector and the ending timepoint is the head of the vector. The boundary conditions are represented by colored splotches. The boundary condition at the beginning of the interval is represented by $x_0$ and the boundary condition at the end of the interval is represented by $x_f$. The BV-AA-MDS solution is the vector colored in black. A BV-AA-prx solution is represented by a curved orange line. For the MS method with $N=3$, IVP’s on the subintervals are represent by blue vectors of varying shades. The magnitude of the residual vector is represented conceptually by the light blue vertical bars. For comparison purposes, an approximate BV-AA-MDS solutions generated by a forward IVP is indicated by a dashed blue vector. The magnitude of the residual vector is represented conceptually by the navy blue vertical bar.
Figure 3.8 Initial MS BV-AA-MDS trajectories, $N = 3$ : mesh 2

BV-AA-MDS solution, approximate BV-AA-MDS solution and BV-AA-prx trajectory

Conceptual diagram of a BV-AA-MDS solution and three approximate BV-AA-MDS. AA-MDS trajectories are indicated by vectors with initial timepoints are the tail of the vector and the ending timepoint is the head of the vector. The boundary conditions are represented by colored splotches. The boundary condition at the beginning of the interval is represented by $x_0$ and the boundary condition at the end of the interval is represented by $x_f$. The BV-AA-MDS solution is the vector colored in black. A BV-AA-prx solution is represented by a curved orange line. For the MS method with $N=3$, IVP’s on the subintervals are represent by blue vectors of varying shades. The magnitude of the residual vector is represented conceptually by the light blue vertical bars. For comparison purposes, an approximate BV-AA-MDS solutions generated by a forward IVP is indicated by a dashed blue vector. The magnitude of the residual vector is represented conceptually by the navy blue vertical bar.
Finding shooting points using a hidden time BV-AA-prx trajectory

Shooting points are desired for the MS method with $N=5$ subintervals. Five IV-AA-MDS trajectories are produced with initial conditions from a hidden time BV-AA-prx trajectory. Weighted RMSD from the initial conformation of the next shooting point is calculated and plotted above for each trajectory. This plot shows that the second assumption described in section 3.4.3.7 is satisfied.
Conceptual diagrams of a nearly converged MS BV-AA-MDS trajectory with 3 subintervals, four restart MS BV-AA-MDS trajectories generated as described above using different rescalings of the initial velocity for a translated forward IVP, and a BV-AA-MDS solution trajectory are shown. AA-MDS trajectories are indicated by vectors with initial timepoints as the tail of the vector and the ending timepoint as the head of the vector. The boundary conditions are represented by colored splotches. The boundary condition at the beginning of the interval is represented by $x_0$ and the boundary condition at the end of the interval is represented by $x_f$. The BV-AA-MDS solution is the vector colored in black. For the MS method with $N=3$, IVP’s on the subintervals are represented by blue vectors of varying shades. The four restart BV-AA-MDS solutions generated by forward IVP’s are indicated by dashed blue vectors.
Figure 3.11  Distribution of total energy in \textit{kcal mol}^{-1}

Distributions of energy within ensemble of trajectories. Trajectories from C\textsubscript{7}\textit{eq} local minimum to C6 local minimum tend to be lower in total energy.
Figure 3.12  Ensemble of trajectories

Ensemble of trajectories. $C_{7\text{eq}} \rightarrow C_6$, $C_{7\text{eq}} \rightarrow C_{5\beta}$, and $C_{7\text{ax}} \rightarrow C_{7\text{eq}}$. 
Figure 3.13  Initial and solution trajectories

Initial and Solution Trajectories. $C_{7\text{eq}} \rightarrow C_{6}$, $C_{7\text{eq}} \rightarrow C_{5\text{ax}}$, and $C_{7\text{ax}} \rightarrow C_{7\text{eq}}$. 
4 SHOOTING METHODS WITH INEXACT BOUNDARY CONDITIONS AND PARAMETER REDUCTION FOR PROTEIN DYNAMICS SIMULATION

4.1 Abstract

Some proteins and other biomolecules are known to make transitions between two known conformations. Assuming the two known conformations can be described adequately, the dynamics of these conformational transitions can be studied mathematically at the atomic level of detail by formulating a boundary value problem (BVP) for ordinary differential equations (ODE’s) and applying a numerical method, such as the multiple shooting method, to find solutions. To apply multiple shooting with a full set of parameters to a BVP with a full set of boundary conditions, as was described in Chapter 3, the number of atoms in the molecule must be limited to avoid excessive computational cost. In this chapter, for the case of single shooting, an alternative simulation approach is presented that involves a reduced set of boundary conditions and a reduced set of parameters. BVP’s are constructed with boundary conditions defined as lower and upper bounds for selected interatomic distances that are intended to approximate potential energy wells. Modeling conformational transitions between potential energy wells has advantages in comparison with modeling conformational transitions between local minima of a potential energy surface. First, the former approach more closely reflects the reality of the physical problem being modeled since biomolecules, in their native environment, are in constant motion, even if the motion is just vibrational and isn’t impacting the overall conformation. Secondly, the former approach allows for the possibility of a reasonably small reduced set of boundary conditions for a large system. Optionally, a boundary condition can also be added to define bounds for the total energy of the system. We also propose an approach for use a reduced parameter set that is based on an application of principles of normal mode analysis. We provide results from the application of these approaches to the study of transitions between potential energy wells for an alanine dipeptide.
4.2 Introduction

Conformational transitions of proteins and other biomolecules play a vital role in critical molecular processes in a healthy living cell or organism. On the other hand, certain conformational transitions of some proteins can cause problems and even lead to diseases (e.g. prion proteins and diseases such as scrapie in sheep, ‘mad cow’ disease in cows, and Creutzfeld-Jacob disease in humans). Assuming the two known conformations can be described adequately, the dynamics of these conformational transitions can be studied mathematically at the atomic level of detail by formulating a boundary value problem (BVP) for ordinary differential equations (ODE’s) and applying a numerical method, such as the multiple shooting (MS) method, to find solutions. This is a boundary value approach to all-atom molecular dynamics simulation (BV-AA-MDS). Successful application of a numerical method produces a solution trajectory, that is, an ordered set of conformations assumed by the molecule at a set of points in time which collectively satisfies a discretized approximation to Newtonian equations of motion and for which the first and last conformations are the aforementioned known conformations. Motivations for all-atom molecular dynamics simulation (AA-MDS), BV-AA-MDS, and use of MS for BV-AA-MDS were provided in Chapter 2 of this dissertation. A strategy for the application of the MS to BV-AA-MDS when the beginning and ending conformations are specified by sets of internal coordinates was proposed in Chapter 3. Also in Chapter 3, the strategy was applied to the study of conformational transitions between local minima of a potential energy surface for an alanine dipeptide.

In section 2.1.4, an AA-MDS trajectory was described mathematically as the solution to either an initial value problem (IVP) or a boundary value problem (BVP). Consider AA-MDS in three dimensions. At any point in time, an atom (indexed by the number \( i \)) is subjected to a force \( \mathbf{f}_i = [f_{i1}; f_{i2}; f_{i3}] \). The motion of the particle for each direction is determined by Newton’s 2\textsuperscript{nd} law, which states that the relationship between an atom’s mass \( m_i \); its acceleration, \( \mathbf{a}_i = [a_{i1}; a_{i2}; a_{i3}] \); and the applied force, \( \mathbf{f}_i \), is \( \mathbf{f}_i = m_i \mathbf{a}_i \). So, there are three 2\textsuperscript{nd}-order scalar differential equations for each atom. We can say, equivalently, that there are six 1\textsuperscript{st}-order scalar differential equations for each atom. For a system with \( n \) atoms, then Newton’s 2\textsuperscript{nd} law...
requires $6n$ 1st-order scalar differential equations. For a two-point boundary value problem, there are boundary conditions consisting of a set of scalar equations that the AA-MDS trajectory must satisfy at two specific points in time. If the number of scalar equations is equal to the number of 1st-order scalar differential equations, then we say here that the set of boundary conditions is a full set of boundary conditions. So, for BV-AA-MDS a full set of boundary conditions consists of $6n$ scalar equations. These scalar equations can be defined as components of a $6n \times 1$ vector valued function, $r$. If there are less than $6n$ boundary conditions, then we say that the BVP has a reduced set of boundary conditions.

Single shooting is a special case of multiple shooting in which there is only one subinterval. In this chapter, we will focus on single shooting for BV-AA-MDS. For one iteration of the single shooting method, a set of initial parameters need to be assigned to generate an AA-MDS trajectory and determine the values of the components of $r$. If the number of parameters is equal to the number of 1st-order scalar differential equations, then we say here that the set of parameters is a full set of parameters. So, for single shooting for BV-AA-MDS, a full set of parameters consists of $6n$ unique parameters assigned at one specific point in time. And, if single shooting is applied with less than $6n$ parameters, then we will say that it is applied with a reduced set of parameters.

(For completeness, we can provide analogous definitions for multiple shooting with $N$. If the number of parameters is equal to the number of subintervals times the numbers of 1st-order scalar differential equations, then we say here that the set of parameters is a full set of parameters. So, for multiple shooting with $N$ subintervals for BV-AA-MDS, a full set of parameters consists of $6nN$ unique parameters with $6n$ parameters assigned at each of $N$ different points in time. We say multiple shooting is applied with a reduced set of parameters when there are less than $6nN$ parameters.)

When the boundary conditions are precisely specified by sets of internal coordinates, as in the BVP’s of Chapter 3, there is a full set of $6n$ boundary conditions. In general, for a BVP with a full set of boundary conditions and for single shooting with a full parameter set, the number of boundary conditions and the number of parameters both scale linearly with the number of atoms in the system. This has
implications from a computational perspective. If the set of parameters on the $k^{th}$ iteration is denoted $s^k$ and the Newton step is denoted by $\xi^N_{\text{newt}(k)}$ and the residual function $F(s)$ is simply the boundary condition function $r$, then the Newton step is computed by the linear system solve

\begin{equation}
F'(s^k)\xi^N_{\text{newt}(k)} = -F(s^k)
\end{equation}

where $F'$ is the $6n \times 6n$ Jacobian of $F$. The computational cost of direct methods for this linear system solve is $O(n^3)$ which becomes prohibitive for large $n$. While general iterative methods for this linear system solve can be considered, it is also reasonable to consider problem-specific alternatives.

In this chapter, we will describe BV-AA-MDS involving a reduced set of boundary conditions and a reduced set of parameters. A conformational transition can be defined as a transition between any pair of conformations that belong to distinct, non-overlapping sets. The sets of conformations that will be defined in this chapter are intended to approximately correspond to wells surrounding local minima of a potential energy surface which we call potential energy wells. The sets will be defined based on lower and upper bounds for selected interatomic distances and will be used to establish boundary conditions. Optionally, a boundary condition can also be added to define bounds for the total energy of the system. It will be argued that these methods allow for the possibility of a reasonably small set of boundary conditions for a relatively large system. We also propose an approach for use a reduced parameter set that is based on an application of principles of normal mode analysis. We provide results from the application of these approaches to the study of transitions between potential energy wells for an alanine dipeptide.

4.3 Preliminaries

4.3.1 All atom molecular dynamics simulation (AA-MDS)

All-atom molecular dynamics simulation (AA-MDS) generally refers to a particular type of molecular modeling in which the motion of the atoms or particles of the molecules of the system are tracked dynamically over a period of time and the motion is governed deterministically by the Newtonian equations of motion. More specifically,
(4.2) \[ M \alpha(t) = f(x(t)), \quad t_0 < t < t_f \]

where \( t \) is a scalar representing time where \( t_0 < t < t_f \); \( x(t), v(t), \) and \( \alpha(t) \) are \( 3n \times 1 \) vectors representing the position, velocity, and acceleration, respectively, of the \( n \) particles of the system at time \( t \) in three dimensions of a rectangular coordinate system; \( M \) is a \( 3n \times 3n \) diagonal matrix with the mass of each particle repeated in three successive diagonal entries; and \( f(x(t)) \) is a \( 3n \times 1 \) vector representing the force acting on each particle of the system at time \( t \) in each dimension. Note that \( v(t) = x'(t) \) and \( \alpha(t) = x''(t) \), so (4.2) is a 2nd-order ordinary differential equation (ODE). In order to begin a simulation, additional specifications are required.

### 4.3.2 Initial value AA-MDS (IV-AA-MDS)

For IV-AA-MDS, additional specifications are the initial values of the form

(4.3) \[ x(t_0) = x_i, \quad v(t_0) = v_i \]

where \( x_i \) and \( v_i \) are \( 3n \times 1 \) vectors and \( t_0 \leq t \leq t_f \). Equations (4.2) and (4.3) define an initial value problem (IVP). For \( f \) linear, the domain for existence and uniqueness of solutions can be specified by inspection of \( f \). For \( f \) nonlinear, the entire domain cannot be specified, but local existence and uniqueness of solutions in an open interval about some \( t = t_i \) can be guaranteed by continuity of \( f \) and \( \partial f / \partial x \) at \( t = t_i \).

### 4.3.3 Boundary value AA-MDS (BV-AA-MDS)

For a two-point boundary value (BV-) AA-MDS, additional specifications are given by

(4.4) \[ r(x(t_0), v(t_0), x(t_f), v(t_f)) = 0 \]

where \( r \) is an \( R \times 1 \) vector for some integer \( R \). In this chapter, we will consider BVP’s with \( R < 6n \). So, the number of scalar boundary conditions in (4.4) is less than the number of scalar differential equations in (4.2).

Equations (4.2) and (4.4) define a two-point boundary value problem (BVP). For a BVP of this form, in general, there may or may not be a solution, and if there is one solution, it might not be the only
one. The adjective ‘two-point’ indicates that \( r \) is a function describing the characteristics of the system at two time-points, \( t_0 \) and \( t_f \). We assume \( f(x) \) is the gradient of a real-valued function, \( U(x) \), called a potential energy function. So, \( f(x) = -\nabla U(x) \).

### 4.3.4 Single shooting (SS)

A description of the multiple shooting method for BV-AA-MDS was provided in section 2.3 based on the presentation in [Asc1995]. This description was based on the assumption that there were \( 6n \) boundary conditions and there were \( 6n \) parameters for each subinterval that determined the initial conditions for the IVP of that subinterval. These assumptions are relaxed in this chapter. We just assume that there are no more than \( 6n \) boundary conditions and no more than \( 6n \) parameters for a subinterval. So, a slightly less cumbersome description is provided here specifically for the single shooting (SS) method in which the number of boundary conditions and the number of parameters are less than or equal to \( 6n \). As was done previously, we rewrite the system of \( 3n \) equations in the 2nd-order ODE (4.2) equivalently as a system of \( 6n \) equations. Then, the BVP (4.2), (4.4) can be written as a 1st-order BVP as given below:

\[
\begin{align*}
(4.5) & \quad y'(t) = h(y(t)), \quad t_0 < t < t_f \\
(4.6) & \quad r(y(t_0), y(t_f)) = 0
\end{align*}
\]

where \( r(y(t_0), y(t_f)) \) is a differentiable function. We will solve the IVP

\[
(4.7) \quad y'(t; s) = h(y(t; s)), \quad y(t_0) = y_0(s), \quad t_0 < t < t_f
\]

where \( s \) is a parameter vector that determines the initial conditions, \( y_0 \). The dependent, but not necessarily equivalent, relationship between \( s \) and \( y_0 \) is a departure from the presentation of multiple shooting in Chapters 2 and 3. The solution for a given \( s \) is \( y(t; s) \) for \( t_0 < t < t_f \). We want to find \( s^* \) such that

\[
(4.8) \quad F(s^*) = r(y_0(s^*), y(t_f; s^*)) = 0.
\]

For any iterative approach with iterates given by \( s_0^*, s_1^*, \ldots, s_k^* \), we may write
\[(4.9) \quad s^{k+1} = s^k + \xi^k.\]

The Newton step, \(\xi^{\text{Nwtn}(k)}\) can be computed by solving
\[(4.10) \quad F'(s^k)\xi^{\text{Nwtn}(k)} = -F(s^k)\]

where \(F'(s)\) is the Jacobian of \(F\) and it takes the form
\[(4.11) \quad F'(s) = B_0 + B_f Y(t)\]

where
\[(4.12) \quad Y(t) \equiv \frac{\partial y(t; s)}{\partial s}\]

and
\[(4.13) \quad B_0 = \frac{\partial}{\partial u} r(u,v), \quad B_f = \frac{\partial}{\partial v} r(u,v).\]

at \(u=\gamma(t_0; s) = s\) and \(v=\gamma(t_f; s)\). Applying Theorem 7.1.8 of [Sto2002], we can find \(Y\) numerically, step-by-step as we solve equation (4.7) by solving the following matrix ODE:
\[(4.14) \quad \frac{d}{dt} Y(t; s) = \frac{\partial h}{\partial y}(y(t; s)) Y(t; s), \quad t_0 < t < t_f\]
\[(4.15) \quad Y(t_0) = I\]

In general, the matrix, \(Y(t)\) will be dense despite the structural sparseness of \(\partial h/\partial y\).

As described in subsection 2.4.3.1, it is expected that the single shooting method will only be effective for BV-AA-MDS for a limited length of the time interval. And, as with multiple shooting, intrinsic error in the numerical methods for solving ODE’s and numerical errors associated with finite arithmetic dictate that the best that can be achieved is an approximate solution to the BVP. The accuracy of the approximation as a Newtonian trajectory and as a solution to the BVP can be expected to depend on the accuracy of the method for numerical solution of IVP’s, mesh selection for IVP solutions, and tolerance selection for determination of a numerical solution to \(F(s)=0\).
4.3.5 All-atom normal mode analysis

Conformational transitions of biomolecules have been studied at the all-atom level using normal mode analysis (e.g. [Bro1983], [Tam2001], [Jaa1998], [Cui2006]). In this subsection, we give a description of all-atom normal mode analysis since it will be used in a method proposed later in this chapter for deriving a reduced set of parameters for single shooting. A 2nd-order Taylor’s series approximation of the potential energy function, \( U(x) \), in an open interval around some position \( x^* \), can be written as

\[
U(x) = U(x^*) + \nabla U(x^*)^T(x - x^*) + \frac{1}{2} (x - x^*)^T \nabla^2 U(x^*)(x - x^*)
\]

Take a look at the right hand side of this equation. The 1st term on the right hand side is a constant. Consider the situation where \( x^* \) is a local minimum. Then, the 2nd term is equal to zero since \( \nabla U(x^*)=0 \).

Since \( f(x) = -\nabla U(x) \), the resulting Taylor approximation of the force at \( x \) is

\[
f(x) \approx -\nabla^2 U(x^*)(x - x^*)
\]

So, the Newtonian equations of motion near a local minimum, \( x^* \), are approximated by

\[
M \ddot{x}(t) = -\nabla^2 U(x^*)(x - x^*)
\]

This is a linear ODE with \( 3n \) independent solutions that can be obtained by determining eigenvectors and eigenvalues of a matrix to be specified below ([Hin2004],[Gol1980]). To obtain the form of these solutions, following [Hin2004], it is useful to transform the system to mass-weighted coordinates, so we let

\[
\dot{x} = M^{1/2} \dot{x} ; \ddot{x} = M^{1/2} \ddot{x} ; \dot{U}(\ddot{x}) = U(x^*) ; \text{and} \quad \nabla^2 \dot{U}(\ddot{x}) = M^{-1/2} \nabla^2 U(x^*) M^{-1/2}
\]

The last equation gives a formula for the mass-weighted Hessian. Then, we can rewrite the above equation as

\[
\dot{x}''(t) = -\nabla^2 \dot{U}(\ddot{x}) (\ddot{x} - \dot{x}^*)
\]

Solutions to this differential equations can be found by determining the eigenvectors and eigenvalues of \( \nabla^2 \dot{U}(\ddot{x}) \). The solutions can be written in the form

\[
\dot{x}(t) = \dot{x}^* + \sum_{i=1}^{3n} \dot{u}_i \cos(\omega_i t + \delta_i)
\]
where \{\omega_i\} are the set of eigenvalues, \{\hat{u}_i\} is a set of eigenvectors of the mass-weighted Hessian at \(x^*\), and \{\delta_i\} is a set of phase factors. The set \{\omega_i\} is also known as the set of vibrational frequencies for the potential energy function at \(x^*\) and \{\hat{u}_i\} is also known as the set of vibrational modes, or normal modes, for the potential energy function. The normal modes specify relatively how far and in what direction each individual atom moves for each of the orthogonal vibrational frequencies. The absolute magnitude is not determined by (4.21) since a set of eigenvectors multiplied by any constant is also a set of eigenvectors.

So, vibrational normal mode analysis classifies all possible motions around a stable equilibrium state by vibrational frequency. There are also \(3n\) independent energy modes which are the eigenvectors of the Hessian of the potential energy function at the corresponding local minimum. Eigenvalues of this Hessian describe the curvature of the potential energy function along the normal mode directions. In other words, these eigenvalues describe the energetic cost of displacing the system by a magnitude of one unit along the direction of the corresponding eigenvector. As described in [Hin2004], there tends to be a significant correlation between energetic modes and vibrational modes. Low-frequency, or slow, vibrational modes are a good approximation to low-frequency energy modes, and vice versa. It could be said that normal mode analysis indirectly classifies possible deformations of a protein from a stable conformation by their energetic costs.

By the above description, we see that there is a theoretical basis for the use of normal mode analysis to study oscillatory motions around a local minimum. Methods based on normal mode analysis which have yet to be justified with a solid theoretical foundation have also been useful for finding feasible pathways between two conformations. These methods, in general, involve following low-frequency modes and calculating normal modes as one moves along a path, e.g., [Bro1983]). They can be understood intuitively as follows: Near a local minimum, collective motions tend to be associated with slow modes and they also tend to be low-energy motions. In contrast, localized motions tend to be associated with faster modes and they tend to be high-energy motions. Larger amplitude motions that are eventually result in conformational transitions tend to occur along the directions of motions that are more collective in nature. These direction tend to be the directions of slower modes ([Hin2004], [Tam2001], [Kun2004], [Son2006]).
4.3.6 Distance measures and distance geometry

In the study of molecular conformations, it is important to have a means for assessing the difference between two conformations. Scalar distance measures can be useful for this purpose. There are various distance measures that have useful applications. In the strategies for solving two-point BVP’s using multiple shooting in Chapter 3, boundary conditions specified by sets of internal coordinates were desired. However, the rectangular coordinate system is convenient to use for AA-MDS. A specific measure, RMSD (Root Mean Squared Deviation), which satisfies some metric properties that are desirable for measurement of distance, was useful as a measurement effectively of similarities in internal coordinates without requiring a transformation from the rectangular coordinate system.

Distance geometry is the characterization and study of sets of points based only on given values of the distances between points belonging to the sets. Applying distance geometry to the study of molecular conformations, the set of points are positions of atoms and the distances between the points are interatomic distances between pairs of atom. There are a variety of useful applications of distance geometry in the study of molecular conformations in fields such as conformational analysis, drug design, NMR spectroscopy, homology modeling, and structure refinement. In some applications, exact distances are not given, but there are lower and upper bounds placed on the distances between pairs of atoms. See [Hav1998], [WuZ2003], or [WuZ2006]) for specific examples. In the realm of distance geometry, one commonly used measure of distance is DME (Distance Matrix Error). For a molecule with atoms indexed from 1 to n and interatomic distances between particles i and j represented by \(d(x_{ij})\), the set of interatomic distances is given by \(D(x) = \{d(x_{ij})\mid i \leq j \leq n\}\). Note \(d(x_{ii}) = 0\) and symmetry of distance measures implies that \(d(x_{ij}) = d(x_{ji})\). If another molecule or another conformation of the molecule has a set of interatomic distances represented by \(D(z) = \{d(z_{ij})\}\), then the distance matrix error between the two conformations is given by

\[
DME(x,z) = \left( \sum_{i=1}^{n} \sum_{j=1}^{n} (x_{ij} - z_{ij})^2 \right)^{1/2}
\]

Other distance measures useful for the study of molecular conformations are defined in [Sch2001].
It is sometimes desirable to classify conformations of a molecule. For example, conformations of a protein are sometimes classified as *folded* or *not folded*. Interatomic distance data can be used in establishing quantitative criteria for classification. One example is the criteria for classification of a β-hairpin fragment as folded that were described in “Atomistic folding simulations of a β-hairpin” in *Journal of Molecular Biology*, 2001 ([Zag2001]). In the next section, we will promote the idea of establishing boundary conditions for BV-AA-MDS using interatomic distances.

### 4.4 Ideas, methods, and analysis

The first three subsections below will be focused on the definition of boundary conditions. The next four subsections will sequentially introduce components of an approach for generated a reduced set of parameters. The next subsection contains a comprehensive strategy for finding solutions to well to well BVP’s. This strategy will be applied the well to well BVP’s for an alanine dipeptide in the next subsection. The last subsection will contain additional analysis and discussion.

#### 4.4.1 Motivation for boundary conditions based on lower and upper bounds for interatomic distances

We consider the prospect of defining boundary conditions based on interatomic distances for use with the single shooting numerical method. For a molecule with $n$ atoms, the complete set of interatomic distances contains $\frac{n(n-1)}{2}$ elements. So, if a scalar boundary condition is included for every interatomic distance of the molecule, then the number of boundary conditions will be greater than the number of rectangular coordinates for the molecule since $\frac{(n^2 - n)}{2} > 3n$ for $n > 3$. So, if a reduced set of boundary conditions is to be attained, then not all interatomic distances can be included. Is there a way of determining how many interatomic distances are necessary for the purpose of adequately defining boundary conditions?

To begin to address this question, we consider another application of interatomic distance data. If the structure of a protein is unknown, nuclear magnetic resonance (NMR) spectroscopy can be used to attempt to determine the structure. Through analysis of NMR data, a distance geometry description of the
molecule can be attained. This description can be expected to include a list of distance constraints -- lower
and upper bounds on the distances between some but not all atom pairs. In this case, the number of
distance constraints necessary to determine the structure can be expected to be a factor of \( n \) not less than 3
([Alt1999]).

For the purpose of defining boundary conditions for single shooting for BV-AA-MDS, not as many distance constraints may be necessary. First, beginning and ending conformations from application of the single shooting can always be inspected. So, if a solution is found using a reduced number of distance constraints and upon closer inspection, the beginning or ending conformation does not belong to the desired potential energy well, then the BVP can be redefined with additional distance constraints added and the single shooting method can be restarted using the solution with an inadequate number of boundary conditions as an initial SS BV-AA-MDS trajectory. So, any inadequacies are self-correcting. Second, an upper bound on total energy level can be added as boundary condition. Within the realm of trajectories with limited total energy, many distance constraints may be redundant. Third, the effective number of boundary conditions for an update on a given iteration can be limited to the active distance constraints -- those which are not already satisfied for the trajectory of the current iteration. Finally, if a good initial SS BV-AA-MDS trajectory is available, then it is likely that the number of active distance constraints will be limited. These ideas will be developed in more detail below.

### 4.4.2 Defining boundary conditions based on lower and upper bounds for interatomic distances

In this subsection, we develop a method for defining boundary conditions in terms of interatomic distances. Assume that the complete set of atom pairs are indexed from 1 to \( n(n - 1)/2 \). And, the beginning and ending structures are specified by bounds on the interatomic distances between sets of index pairs of size \( \sigma \) and \( \varsigma \), respectively where

\[
[b_1; b_2; \ldots; b_\sigma]
\]

\[
[e_1; e_2; \ldots; e_\varsigma]
\]
and each $b_k$ and $e_k$ correspond to an atom pair for which there is a distance constraint for the starting and ending conformations, respectively. Let

\begin{equation}
(4.25) \quad l = [l_{b_1}; l_{b_2}; \ldots; l_{b_\sigma}; l_{e_1}; l_{e_2}; \ldots; l_{e_\varsigma}] \]
\end{equation}

\begin{equation}
(4.26) \quad u = [u_{b_1}; u_{b_2}; \ldots; u_{b_\sigma}; u_{e_1}; u_{e_2}; \ldots; u_{e_\varsigma}] \]
\end{equation}

\begin{equation}
(4.27) \quad d = [d_{b_1}; d_{b_2}; \ldots; d_{b_\sigma}; d_{e_1}; d_{e_2}; \ldots; d_{e_\varsigma}] \]
\end{equation}

be $\sigma+\varsigma \times l$ vectors of the lower bounds, upper bounds and actual values for the distances of the $\sigma+\varsigma$ selected atom pairs for the beginning and ending conformations. Then, for $i(i) \equiv [b_1; b_2; \ldots; b_\sigma; e_1; e_2; \ldots; e_\varsigma]$, the $i^{th}$ component of the nonlinear boundary condition function, $r = \{r_i(y(t_0), y(t_f))\}$, is given by

\begin{equation}
(4.28) \quad r_i(y(t_0), y(t_f)) = \text{MAX}((d_i - u_i), (l_i - d_i))
\end{equation}

The function defined by the notation $(\cdot)^+$ evaluates to the maximum of zero and $\cdot$. For vector arguments, let the $\text{MAX}(\cdot, \cdot)$ function and the $(\cdot)^+$ function be defined componentwise. Then we may write

\begin{equation}
(4.29) \quad r(y(t_0), y(t_f)) = \text{MAX}((d - u), (l - d))
\end{equation}

If boundary conditions are intended to represent a potential energy well, one way to establish the lower and upper bounds is by performing IV-AA-MDS simulations with an initial position corresponding to the primary local minimum of the well and various initial momenta small in magnitude. The lower and upper bounds for interatomic distances from portions of these samples for which the molecule remains in the potential energy well can then be recorded. The Jacobian matrices of (4.13) can be computed analytically by calculations similar to those used for computing the Hessian for the van der Waals components of a biomolecular potential energy function.

### 4.4.3 Defining a boundary condition for an upper bound on total energy

For an isolated system, the Hamiltonian, $H$, is constant in theory. Ignoring the numerical error and discretization error in the calculation of $H$ that are introduced as the simulation evolves, one can specify an
upper bound for $H$ throughout a simulation by specifying an upper bound for $H$ at a single point in time. So, for $H$ defined as

$$
H(x,m,v) = E_{\text{TOTAL}}(x, m, v) = U(x(t)) + \frac{1}{2} \sum_{i=1}^{n} \left\| v_{3i-2}(t) \right\|_2^2 + \left\| v_{3i-1}(t) \right\|_2^2 + \left\| v_{3i}(t) \right\|_2^2
$$

a constraint can be specified by adding a component to $r = \{ r_i(y(t_0), y(t_f)) \}$, which puts a constraint on the initial conditions of the form

$$
(4.31) \quad r_i(y(t_0), y(t_f)) = (H(x(t_0), m, v(t_0)) - H_u).
$$

where $H_u$ is an upper bound on the total energy of the system. Adding this component to $r$ requires an additional row in the Jacobian matrices of (4.13), but the entries of this row are fairly easy to compute analytically.

### 4.4.4 Fixing a subset of initial parameters to satisfy initial boundary conditions of separable BVP

A BVP of the form (4.5), (4.6) has separable boundary conditions if the function $r$ of (4.6) can be written in the form

$$
(4.32) \quad r(y(t_0), y(t_f)) = [ r_0(y(t_0)); r_f(y(t_f)) ]
$$

where $r_0$ and $r_f$ and $\sigma \times I$ and $\varsigma \times I$ are vector-valued functions and $\sigma + \varsigma = \mathbb{R}$. In the next subsection, we will describe a strategy for generating a reduced set of parameters, by use of a normal mode selection approach, that we hypothesize will retain much of the utility of a full set of parameters. The presentation and initial evaluation of this approach is simplified by considering first a BVP of the form (4.5), (4.6) with separable boundary conditions with $R/2$ boundary conditions on $x(t_0)$ and $R/2$ boundary conditions on $x(t_f)$. In terms of the residual function, $F$ of (4.8), write $F = [ F_0; F_f ]$ where $F_0$ and $F_f$ are components of $F$ corresponding to the boundary conditions on $x(t_0)$ and $x(t_f)$, respectively. Consider single shooting with a full set of $6n$ parameters corresponding to $3n$ rectangular coordinates for $x(t_0)$ and $3n$ velocities in rectangular...
coordinates at \( x(t_0) \). In terms of the parameter vector, \( s^0 \) for (4.7), write \( s^0 = [x^0; v^0] \) where \( x^0 \) and \( v^0 \) are initial position and initial velocity vectors for the initial iteration of the single shooting method. It is usually trivial to choose \( 3n \) rectangular coordinates for \( x(t_0) \) to satisfy the \( R/2 \) boundary conditions on \( x(t_0) \), and, in any case, a numerical method for solving ODE’s is not required for this purpose. So, assume that \( F_0(s^0) = 0 \). Further, assume that a global convergence scheme for updating the parameter satisfies, \( s^k = [x^0; v^k] \) for all \( k \). Then, \( F_0((s^k)) = F_0(s^0) = 0 \), for all \( k \). So, the BVP is reduced finding a \( 3n \) initial velocity vector \( v^k \) that satisfies \( F_f(s^k) = 0 \). The only element of the Jacobian that are relevant for this problem are the entries in the \( 3n \) columns corresponding to the initial velocities and the \( R/2 \) rows corresponding to the components of \( F_f \). So, there are \( R/2 \) equations and \( 3n \) parameters. So, the equation

\[
(4.33) \quad F_f'(v^k) \xi^{k}_v = -F_f(v^k)
\]

effectively replaces equation (4.10) where \( v^k \) is written as the argument to \( F_f \) to emphasize that the \( v^k \) block is the only block that gets updated, \( F_f'(v^k) \) is the \( R/2 \times 3n \) reduced Jacobian described above, and \( \xi^{k}_v \) is the step from finding a solution or approximate solution to this system of scalar equations with \( 3n \) unknowns and \( R/2 \) equations. The method described in this section will be referred to as **separable boundary condition dimension reduction**. We will discuss solving the system of equations (4.33) in further detail in a later subsection.

### 4.4.5 Parameterization using normal mode directions

Consider the separable BVP described in the previous subsection in which there is a local minimum of the potential energy surface, \( x_0 \), such that setting \( s^0 = [x_0^0; v^0] \) results in \( F_0(s^0) = 0 \). Assume that this local minimum has mass-weighted Hessian, \( \nabla^2 U(\hat{x}_0) \), with \( n \) linearly independent eigenvectors. Now, apply the separable boundary condition dimension reduction method with \( s^k = [x^0_0; v^k] \) for all \( k \). For single shooting, there are \( 3n \) parameters—the \( 3n \) components of \( v^k \) to iteratively determine. Each component corresponds to a direction of one atom in one of the three rectangular coordinate directions. This set of directions is linearly independent. We can consider a different set of \( 3n \) parameters—initial velocities in the \( 3n \) directions of a set, \( \{\hat{u}_i\} \), of orthogonal eigenvectors of the symmetric matrix \( \nabla^2 U(\hat{x}_0) \). Denote a parameter
vector defined this way by \( v_u^k \). We have \( F'_f(v_u^k) = F'_f(v^k) \hat{u} \) where \( \hat{u} \) is a \( 3n \times 3n \) matrix of the eigenvectors of \( \nabla^2 \hat{U}(\hat{x}_0) \). Then, (4.33) is replaced by

\[
(4.34) \quad F'_f(v_u^k) \tilde{\xi}_v^{u(k)} = -F'_f(v_u^k)
\]

where the components of \( \tilde{\xi}_v^{u(k)} \) are the updates to the velocities in the direction of the eigenvectors of \( \nabla^2 \hat{U}(\hat{x}_0) \). To obtain the resulting velocities in the rectangular coordinate directions the formula

\[
(4.35) \quad \hat{\xi}_v^k = \hat{u} \tilde{\xi}_v^{u(k)}
\]

may be applied. Note that (4.34) implies that

\[
(4.36) \quad \xi_v^{u(k)} = -\left( F'_f(v_u^k) \right)^{-1} F'_f(v_u^k) = -\left( F'_f(v^k) \right)^{-1} F'_f(v_u^k) = \hat{u}^{-1} \hat{\xi}_v^k
\]

So, \( \xi_v^k = \hat{u} \tilde{\xi}_v^{u(k)} = \hat{\xi}_v^k \). Therefore, while conceptually this approach might seem different, the end result is the same parameter update.

### 4.4.6 Parameterization using a subset of normal mode directions

The separable boundary condition dimension reduction method effectively reduces the full set of single shooting parameters from \( 6n \) to \( 3n \), since the method determines fixed values for \( 3n \) parameters prior to applying single shooting. It is still of interest to reduce the number of parameters further. While this could be done in rectangular coordinates by choosing to fix the initial velocities of some particles in certain rectangular coordinate directions. But, there is a way to choose a reduced set of parameters that seems more appropriate from a theoretical standpoint. Recall that near a local minimum, larger amplitude motions and collective motions that eventually result in conformational transitions of a molecule tend to occur along the directions of slower modes. If the solution to a boundary value problem is hypothesized to require a conformation transition corresponding to a global or collective motion of a molecule, then it makes sense to consider a reduced set of parameters corresponding to adjustments to initial velocities in the direction of the slowest modes, that is in the direction of eigenvectors with the smallest non-zero eigenvalues. Modification of the initial velocities in these directions may have a more significant effect relative to modification of initial velocities in directions of faster modes. So, if \( \hat{u} \) is a matrix with \( \hat{N} \)
columns selected as a subset of the columns of \( \dot{u} \). Then, (4.34), a system of \( R/2 \) scalar equations and \( 3n \) scalar unknowns is replaced by

\[
(4.37) \quad F_f'(v_u^k)\xi_v^{(k)} = -F_f(v_u^k),
\]

a system of \( R/2 \) scalar equations and \( \tilde{N} \) scalar unknowns. So, the reduced set of parameters contains \( \tilde{N} \) parameters instead of \( 3n \). To determine the initial velocities in the rectangular coordinates, the formula

\[
(4.38) \quad \xi_v^{(k)} = \tilde{\xi}_v^{(k)}
\]

may be applied. For \( \tilde{N} < 3n \), in general, \( \xi_v^{(k)} \neq \tilde{\xi}_v^{(k)} \).

### 4.4.7 Parameter update methods

In equation (4.37), an \( R/2 \times \tilde{N} \) matrix multiplied by an \( \tilde{N} \times 1 \) unknown parameter vector that must equal an \( R/2 \times 1 \) vector. We will assume that the \( R/2 \times \tilde{N} \) matrix is well-conditioned and of full rank. If it isn’t well-conditioned, the time interval of the BVP may be too long for single shooting. So, by definition of full rank, \( \text{RANK}(F_f'(v_u^k)) = \text{MAX}(R/2,\tilde{N}) \). If \( R/2 = \tilde{N} \), there is a unique solution that can be obtained (for example, by Gaussian elimination). Consider \( R/2 \neq \tilde{N} \). One popular approach to solving the system of equations in this situation is to compute the Moore-Penrose inverse of \( F_f'(v_u^k) \) which we denote by \( F_f'(v_u^k)^{-} \) ([Sto2002]). The Moore-Penrose inverse, \( B^+ \), of a matrix, \( B \), satisfies the following four properties ([Sto2002]).

1. \( BB^+B = B \)
2. \( B^+BB^+ = B^+ \)
3. \( (BB^+)^T = BB^+ \)
4. \( (B^+B)^T = B^+B \)

If \( (B^TB)^{-1} \) exists, then \( B^+ = (B^TB)^{-1}B \). Assuming \( F_f'(v_u^k) \) is of full rank, we can apply the Moore-Penrose inverse to (4.37) to get

\[
(4.39) \quad \xi_v^{(k)} = -F_f'(v_u^k)^{-} F_f(v_u^k)
\]
If \( R/2 > \tilde{N} \), the system is overdetermined and

\[
F_f'(v^k_u) = \left( F_f'(v^k_u)^T F_f'(v^k_u) \right)^{-1} F_f'(v^k_u)^T
\]

Substitution into (4.39) reveals that this is equivalent to a least squares solution. And if \( R/2 < \tilde{N} \), the system is underdetermined and

\[
F_f'(v^k_u) = F_f'(v^k_u)^T \left( F_f'(v^k_u) F_f'(v^k_u) \right)^{-1}
\]

An alternative to the Moore-Penrose solution can be derived by defining a modified residual function, with \( \tilde{N} \) components, as

\[
F_f^\tilde{u} (v^k_u) = \bar{u}^T [F_f(v^k_u)...(0;...;0)]
\]

where additional components with values of zero may need to be added if \( R/2 < 3n \). Then, (4.37) is replaced by

\[
F_f^{\tilde{u}} (v^k_u) \mathbf{z}^{\tilde{u}(k)} = -F_f^\tilde{u} (v^k_u),
\]

which is a system of \( \tilde{N} \) equations and \( \tilde{N} \) unknowns and therefore has a unique solution as long as the matrix is non-singular. If \( F_f^\tilde{u} (v^k_u) \) is zero, then the original residual function, \( F_f (v^k_u) \) is orthogonal to each column vector of \( \bar{u} \). So, the residual is orthogonal to the column space of \( \bar{u} \). While it may be that

\[
F_f(v^k_u) \neq 0, \ v^k_u \text{ is a minimizer of } \|F_f^\tilde{u} (v^k_u)\|^2 \text{ for all } v \text{ in the column space of } \bar{u}.
\]

This method has some promise if the normal mode directions that are not selected are not significant for the conformational transition. If the normal mode directions that are not selected are not significant, it could be that \( F_f(v^k_u) = 0 \) even though the modified function doesn’t specifically specify this requirement if not all normal modes are selected. This alternate approach is less conventional and is more dependent on the sufficiency of the selected normal mode directions. It makes sense that a subset of normal mode directions might be sufficient for initial conditions. But, this use of the subset in a modified residual function carries an implied assumption that these directions are also important for ending conditions. This is a questionable assumption, so this alternative approach will need to undergo some
testing and analysis to see if it is appropriate for general use. It may have some appeal for particular
definitions of $F_j(v^k_u)$. This will be discussed further in another subsection.

### 4.4.8 A SS BV-AA-MDS strategy for well$\rightarrow$well transitions

Here we provide a strategy for defining and finding solutions to well to well BVP’s. The problem is as follows.

**Problem:**

For a given molecule with $n$ atoms and a corresponding potential energy function, suppose that internal coordinates of two different local minima for the potential energy function have been identified. Define a BVP by determining a beginning and ending time and determining boundary conditions such that they approximate potential energy wells corresponding to the two local minima. Then apply a single shooting method to find a solution to the BVP, subject to the following conditions:

1. There are no more than $R$ scalar boundary conditions where $0 < R < 3n$.
2. There are no more than $\hat{N}$ parameters where $0 < \hat{N} \leq 3n$.
3. The total energy of the system is less than $H_u$.

The solution to the BVP will be a trajectory that satisfies the Newtonian equations of motion and for which the molecule transitions from one potential energy well to the other. The initial time, $t_0$, may be set to zero. This affords no loss of generality since the right hand side of (4.2) is not an explicit function of $t$. So, let $t_0 = 0$. Let coordinates in some rectangular coordinate system of the beginning and ending local minima with the internal coordinates described above be represented by $x_0$ and $x_f$.

**A SS BV-AA-MDS strategy for well$\rightarrow$well transitions:**

1. **Formulate as BVP.** Formulate a BVP with
a. the separable boundary condition dimension reduction method applied with the initial position set to the local minimum of the beginning potential well.

b. boundary conditions for beginning and ending conformations defined in terms of lower and upper bounds for interatomic distances of selected atom pairs.

c. a boundary condition for an upper bound on total energy.

So, using the notation of the previous subsections, the BVP will take the form

\[ M a(t) = f(x(t)), \ 0 < t < t_f \]

\[ x(0) = x_0 \]

\[ l = [l_e_1, l_e_2, \ldots, l_e_\varsigma], u = [u_e_1, u_e_2, \ldots, u_e_\varsigma], d = [d_e_1, d_e_2, \ldots, d_e_\varsigma] \]

\[ r_f(x(t_f)) = \max((d - u)_+, (l - d)_-) \]

\[ r_f(x(0), v(0)) = (H(x(t_0), m, v(t_0)) - H_0) \]

\[ r(x(0), v(t_f)) = [r_f(x(t_f)); r_f(x(0), v(t_f))] = 0 \]

where \( r_f \) is an \( \varsigma \times 1 \) vector-valued function, so \( r \) has \( R = \varsigma + 1 \) components. Assuming the ending time and the selected atom pairs and the corresponding bounds have not already been obtained, the next three steps below provide an approach for obtaining them.

2. **Develop criteria to identify conformations in ending potential energy well.** Develop criteria for the identification of conformations in the ending potential energy well. These criteria may be defined in terms of important dihedral angles, potential energy levels, global distance measures like RMSD or DME, or possibly interatomic distances for a small subset of atom pairs. The latter could correspond to specific hydrogen bonds or native contacts. Hydrogen bonds are a type of attractive interaction that exists between an electronegative atom and a hydrogen atom bonded to another electronegative atom. Two atoms are generally
considered to be in contact if they are within a specified distance of each other. A native contact is a contact that is characteristic of a particular conformation of the molecule.

3. **Select subset of atom pairs.** If an adequately sized subset of atom pairs was not determined in the previous step, select a subset of atom pairs which adequately define the ending conformation. The subset selection may include only those atoms pairs corresponding to native contacts in one of the conformations. They could also be selected based on atom type (e.g. only atom pairs in which both are heavy atoms). The former approach would involve selection of only those pairs for the interatomic distances that are less than or equal to some minimum value for a specified proportion of the simulations in the corresponding well.

4. **Determine lower and upper bounds for interatomic distances.** For the ending potential well, perform a series of initial value simulations with an initial position corresponding to the local minimum and with initial velocities of varying magnitude ranging from arbitrarily small to a magnitude large enough to observe an escape from the potential energy well. Identify the portions of the simulations for which the molecule is in the two potential energy wells of interest. Record the lower and upper bounds for these selected atom pairs and use them to define boundary conditions for each potential energy well by assigning values to the components of the functions $l$ and $u$ as defined in subsection 4.4.2.

5. **Obtain subset of eigenvectors of mass-weighted Hessian.** The reduced set of parameters that is prescribed by item 2 of the problem statement can be accomplished by defining the parameters to be the multiples of a subset of the $3n$ eigenvectors of the mass-weighted Hessian of the potential energy function, $\nabla^2 \hat{U}(\hat{x}^*) = M^{-1/2} \nabla^2 U(x^*) M^{-1/2}$ such that the multiples a determine a linear combination of these vectors which define $v(\theta)$. Let the subset be the column vectors of the $3n \times \tilde{N}$ matrix, $\tilde{u}$. The subset can be chosen based on an assumption about the likelihood that there is a trajectory with an initial velocity chosen in the span of the column space of $\tilde{u}$ that accomplishes the desired conformational transition.
6. **Find an approximate BV-AA-MDS trajectory.** Find an approximate BV-AA-MDS trajectory with an initial velocity in the span of the column space of $\dot{u}$. Use this approximate solution to assign $t_f$. This type of trajectory and the assignment of $t_f$ can possibly be generated using the strategies similar to those outlined in section 3.4.3.2. With the added inequality constraint for total energy, there are different types of approximate BV-AA-MDS trajectories that could be used. First, a trajectory may be sufficiently low in total energy but may not transition to the ending potential well. Second, a trajectory may move to the ending potential energy well, but the total energy criterion may not be met. Third, a trajectory may not meet either of the criteria.

7. **Apply the single shooting algorithm.** Apply the single shooting algorithm using the initial velocities of the approximate BV-AA-MDS approximate trajectory identified above as an initial guess for the solution. The parameter update for each may be accomplished using one the methods described in section 4.4.7 along with a global convergence scheme like those described in 2.3.6. All of the scalar boundary conditions that comprise the function $r$ are satisfied on intervals. The precise location of the endpoints of the intervals for which these boundary conditions are satisfied, in reality, were not developed with a great amount of precision. So, if a few of the boundary conditions are not quite satisfied, the result may still be, in practical terms, a solution trajectory. With this in mind, on a given iteration, for any boundary condition that is satisfied, we consider these to be inactive boundary conditions. During the parameter update, these boundary conditions can be ignored by simply removing the appropriate rows of $F$ and $F'$. So, on a given iteration, the number of active boundary conditions could be anywhere between $0$ and $R$. 
4.4.9 Applying the SS strategy to study of conformational transitions of 
*N-acetyl-N'-methylalaninamide*

The twenty-two atom dipeptide *N-acetyl-N'-methylalaninamide* and a potential energy function that we has been used in the study of this molecule were introduced in earlier in this dissertation. Here we study this alanine dipeptide in *vacuo* to provide an illustration of the use of the strategy outlined in the previous subsection. To assess the conformation of the alanine dipeptide, the C-N-C\(^\alpha\)-C dihedral angle (\(\phi\)) and the N-C\(^\alpha\)-C-N dihedral angle (\(\psi\)) of the alanine residue are of primary importance. A projection onto a two-dimensional subspace determined by the values of \(\phi\) and \(\psi\) will again be used as a way to visualize the potential energy surface and also as a way to visualize conformational change.

**Problem:**

The problem is to obtain trajectories corresponding to conformational transitions from the C7\(_{ax}\) potential energy well to the C7\(_{eq}\) potential energy well, i.e. C7\(_{ax}\) well \(\rightarrow\) C7\(_{eq}\) well.

1. **Formulate as BVP.** With respect to the problem description of the previous subsection, we consider several different sets of additional conditions, formulating 27 unique BVP’s. In all of the problems, the maximum number of boundary conditions is \(R = 26\) for all members of each set. The sets can be grouped into three subsets based on the upper limit on total energy; \(H_u\) is set to 40, 0 and \(-10\) for sets 1, 2, and 3, respectively. Each set has nine different specifications on the number of parameters. \(\tilde{N}\) is set to 66, 30, 20, 10, 9, 8, 7, 6, and 5. Let coordinates in some rectangular coordinate system of the primary local minimum of a particular C7\(_{ax}\) potential energy well and a particular C7\(_{eq}\) potential energy well be represented by \(x_0\) and \(x_f\).

2. **Develop criteria to identify conformations of potential energy wells.** For the identification of conformations of the ending potential energy well, we use the values of the \(\phi\) and \(\psi\) angles. The criteria are depicted visually in Figure 4.1. This is \(\phi-\psi\) plot in which the values for the \(\phi\)
and $\psi$ angle pairs which satisfy the criteria are in the region enclosed by the thick black contour line.

3. **Select subset of atom pairs.** For the selection of atom pairs alanine dipeptide, we select all interatomic distances between heavy atoms separated by at least two atoms on the peptide chain.

4. **Determine lower and upper bounds for interatomic distances.** A series of nine IV-AA-MDS simulations were performed with an initial position beginning in the C7$_{eq}$ local minimum. Initial directions were chosen randomly from a Gaussian distribution. Each IVP in the series had a duration of 12.58 AKMA units, or about 0.6 ps. Values of total energy in kcal mol$^{-1}$ for the 9 simulations were $-20$, $-19$, $-17$, $-15$, $-13$, $-8$, $-3$, $10$, and $20$. By inspection of $\phi$-$\psi$ plots, the sizable sample of the portion of the simulations for which the molecule was approximately in the potential energy well was retained. $\phi$-$\psi$ plots corresponding to these portions of the simulations are given in Figure 4.2(a). For comparisons to be discussed later, an identical analysis was done for the C7$_{ax}$ potential energy well, and a corresponding plot is shown in Figure 4.2(b). In Figure 4.3, the distribution of the 25 selected heavy atom distance ranges for the two conformation types are shown. These ranges were used to assign values to the vector of upper and lower bounds (i.e. $l$ and $u$). The indices used in Figure 4.3 are provided in Table 4.1.

5. **Obtain subset of eigenvectors of mass-weighted Hessian.** For this BVP, the conformational transition from C7$_{ax}$ to C7$_{eq}$ is assumed to be a global, coordinated, and collective motion of the molecule. In terms of normal modes, this assumption suggests selection of slow normal modes, that is, normal modes with small eigenvalues. So, for each of the values of $\tilde{N}$ which specify a different number of parameters, the eigenvectors for $\tilde{N}$ smallest eigenvalues were selected as the set of directions.
6. **Find an approximate BV-AA-MDS trajectory.** For each BVP, a different approximate BV-AA-MDS trajectory was used as an initial SS BV-AA-MDS trajectory. The three initial trajectories with $\tilde{N}=66$ correspond to BVP’s with a full set of parameters. The total energy of the system for these three initial trajectories were 40, 15, and 40 kcal mol$^{-1}$, respectively. For the other values for $\tilde{N}$, the initial trajectories were generated by only including the components of the initial velocities for these three trajectories in the directions of the $\tilde{N}$ slowest modes. For all BVP’s the ending time, $t_f$, was set to $t_f = 189(0.015724) \text{AKMA units} \approx 0.15 \text{ps}$.

7. **Apply the single shooting algorithm.** The single shooting algorithm was applied using the latter of the two parameter update method described in subsection 4.4.7. A trust-region dogleg global convergence scheme (see subsection 2.3.6) was used with a maximum of 26 iterations. The convergence results are summarized in Table 4.3. The number of iterations is given as well as the initial and final total energy and the initial and final values for the mass-weighted objective function, $F$. In Figure 4.4, $\phi-\psi$ overlay plots are shown for the final trajectories for all three sets of BVP’s. Figure 4.5, Figure 4.6, Figure 4.7, and Figure 4.8 serve to illustrate the difference between the initial and ending trajectories on a $\phi-\psi$ plot for the all-mode selection (i.e. $\tilde{N} = 66$) and for the selection of the five slowest modes. Figure 4.9 shows a distribution of the total energy for the final trajectories by the three sets. For the twenty-two atoms of alanine dipeptide, there are $22 \times 3 = 66$ vibrational modes. Six of the sixty-six vibrational modes correspond to rotational and translational motion and the frequencies of these six vibrational modes are all 0. A profile of the frequencies (in wavenumbers) of the other 60 vibrational modes is provided in Figure 4.10.
4.4.10 Additional analysis of \textit{N}-acetyl-N'\texttext{-methy}lalaninamide study and further discussion

Below are some additional comments regarding the study of the alanine dipeptide and its implications.

1. Results reported in [Tam2001] seem relevant. In this work, vibrational modes were not used as initial velocities for AA-MDS as we have done here, but, from a more coarse-grained analysis, it was asserted that often one vibrational mode is closely correlated with a particular conformational change. In each of the three sets of BVP’s described above, the number of parameters, \( \tilde{N} \), is set to different values — 66, 30, 20, 10, 9, 8, 7, 6, and 5 — to solve the BVP’s. The parameters were chosen as set of \( \tilde{N} \) vibrational modes of the lowest frequency. Simulations were also performed for \( \tilde{N} = 2, 3, \) and 4 and for \( \tilde{N} = 1 \) with the one parameter being one of the slowest three modes. None of the these simulations resulted in solution trajectories. Still, it could be that there using only one parameter (or, at least, less than five would result in solution trajectories, if appropriate modes were selected). This would be worth examining more closely in the future.

2. Consider the normal mode direction selection strategy. As a sort of control, for \( \tilde{N} = 10 \) and 5, the strategy described above was modified to select the \( \tilde{N} \) fastest modes. These simulations terminated unsuccessfully with trajectories exhibiting high values for total energy and scattered \( \phi-\psi \) paths that did not lead to the \( \text{C7}_e\) potential well. These results providing additional evidence for the usefulness of the slow mode selection strategy.

3. It would be worthwhile to compare of the results presented here with results from simulations performed using the Moore-Penrose parameter update method.

4. The results presented here include 26 boundary conditions—a reduction from the full set of \( 3n=66 \) boundary conditions. It would be interesting to attempt a further reduction. In particular the atom pairs might be further reduced to just include those for which the range of interatomic distances
for the two potential energy wells are disjoint or almost disjoint. In this case, inspection of Figure 4.3 suggests the number of atom pairs of this type is only about five.

5. For the trust region global convergence scheme, scaling can affect convergence. In particular, it is important to consider appropriate relative scaling of the boundary conditions for distance constraints and the boundary condition for the total energy constraint. Scaling was applied with the intent that for a typical trajectory that is not a solution, that the components of $r$ would be approximately the same, as is prescribed in [Den1996].

6. We infer from the summary data in Table 4.3, Figure 4.4, Figure 4.5, Figure 4.6, Figure 4.7, and Figure 4.8 that using a smaller number of parameters tended to give more rapid convergence to solutions that were lower in total energy. In future work, we plan to investigate these trends for a larger data set and bigger systems.

7. It would be worthwhile to investigate more thoroughly the dependence between convergence and the choice for the initial trajectories.

8. Figure 4.9 shows a relationship between the energy constraint and the energy of the ending solution. For BVP set 1, the maximum energy constraint of $40 \text{kcal mol}^{-1}$ is easily achieved and was not an active constraint in most cases. For BVP set 2, the maximum energy constraint of $0 \text{kcal mol}^{-1}$ seems achievable in most cases, but not so easily perhaps, as it is a severe constraint for most cases. For BVP set 3, the maximum energy constraint of $-10 \text{kcal mol}^{-1}$ seems to perhaps not be achievable based on the observed results. But, using this constraint results in trajectories of lowest energy in comparison with the other two sets.

## 4.5 Summary

We have proposed the application of single shooting methods – a numerical method for solving boundary value problems for ordinary differential equations – to find all-atom molecular dynamics trajectories corresponding to conformational transitions in proteins from one well of a potential energy
surface to another. We have successfully applied this approach to find transitions between C7_{ax} and C7_{eq} potential energy wells for the alanine dipeptide, \textit{N-acetyl-N'}-methylalaninamide. A reduced number of boundary conditions is defined based on the requirement that a set of bounds for interatomic distances between non-adjacent heavy atoms are satisfied for the beginning and ending structures. The set of ranges is derived empirically by performing low energy simulations in each potential energy well and analyzing the range of interatomic distances. As we are often more interested in an ensemble of trajectories that move between potential energy wells rather than trajectories that move precisely between local minima, the definition of boundary conditions presented here is useful in the sense that it can be an accurate way to represent the transition of interest. Also, it may be extended conveniently to larger molecules. We have also presented a normal-mode-based approach for reducing the number of parameters used in applying single shooting to solving these well to well BVP problems, and this approach has also been successfully applied for finding transitions between the C7_{ax} and C7_{eq} potential energy wells for the alanine dipeptide. The computational approach developed here applies only to single shooting. It would be worthwhile to consider a similar computational approach that is applicable for multiple shooting with an arbitrary number of subintervals.

4.6 References


Figures and Tables

Table 4.1  Indices of atom pairs for N-acetyl-N'-methylalaninamide

<table>
<thead>
<tr>
<th>Index</th>
<th>Pair ID</th>
<th>Atom 1</th>
<th>Atom 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2 9</td>
<td>ACE; CH3</td>
<td>ALA; CA</td>
</tr>
<tr>
<td>2</td>
<td>2 11</td>
<td>ALA; CB</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2 15</td>
<td>ALA; C</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2 16</td>
<td>ALA; O</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2 17</td>
<td>NME; N</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>2 19</td>
<td>NME; CH3</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>5 11</td>
<td>ACE; C</td>
<td>ALA; CB</td>
</tr>
<tr>
<td>8</td>
<td>5 15</td>
<td>ALA; C</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>5 16</td>
<td>ALA; O</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>5 17</td>
<td>NME; N</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>5 19</td>
<td>NME; CH3</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>6 9</td>
<td>ACE; O</td>
<td>ALA; CA</td>
</tr>
<tr>
<td>13</td>
<td>6 11</td>
<td>ALA; CB</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>6 15</td>
<td>ALA; C</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>6 16</td>
<td>ALA; O</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>6 17</td>
<td>NME; N</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>6 19</td>
<td>NME; CH3</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>7 16</td>
<td>ALA; N</td>
<td>ALA; O</td>
</tr>
<tr>
<td>19</td>
<td>7 17</td>
<td>NME; N</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>7 19</td>
<td>NME; CH3</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>9 19</td>
<td>ALA; CA</td>
<td>NME; CH3</td>
</tr>
<tr>
<td>22</td>
<td>11 16</td>
<td>ALA; CB</td>
<td>ALA; O</td>
</tr>
<tr>
<td>23</td>
<td>11 17</td>
<td>NME; N</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>11 19</td>
<td>NME; CH3</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>16 19</td>
<td>ALA; O</td>
<td>NME; CH3</td>
</tr>
</tbody>
</table>

Indices of heavy atom pairs separated by two or more atoms on the chain for N-acetyl-N'-methylalaninamide. Heavy atoms are labeled in the figure below.
Table 4.2  Classification of atom pairs for N-acetyl-N′-methylalaninamide
Separation on chain and atom type

<table>
<thead>
<tr>
<th>atom id</th>
<th>residue id</th>
<th>mass</th>
<th>atom type</th>
<th>id</th>
<th>atom pair includes at least one hydrogen atom.</th>
<th>heavy atom pair with less than two heavy atoms between them.</th>
<th>heavy atom pair with exactly two heavy atoms between them.</th>
<th>heavy atom pair with more than two heavy atoms between them.</th>
</tr>
</thead>
<tbody>
<tr>
<td>HH13</td>
<td>ACE</td>
<td>1.00800002</td>
<td>H</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CH3</td>
<td>ACE</td>
<td>12.0109997</td>
<td>C</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HY12</td>
<td>ACE</td>
<td>1.00800002</td>
<td>H</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HY13</td>
<td>ACE</td>
<td>1.00800002</td>
<td>H</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>ACE</td>
<td>12.0109997</td>
<td>C</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O</td>
<td>ACE</td>
<td>15.9989996</td>
<td>O</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>ALA</td>
<td>14.0107</td>
<td>N</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>ALA</td>
<td>1.00800002</td>
<td>H</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CA</td>
<td>ALA</td>
<td>12.0109997</td>
<td>C</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HA</td>
<td>ALA</td>
<td>1.00800002</td>
<td>H</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CB</td>
<td>ALA</td>
<td>12.0109997</td>
<td>C</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HB1</td>
<td>ALA</td>
<td>1.00800002</td>
<td>H</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HB2</td>
<td>ALA</td>
<td>1.00800002</td>
<td>H</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HB3</td>
<td>ALA</td>
<td>1.00800002</td>
<td>H</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>ALA</td>
<td>12.0109997</td>
<td>C</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O</td>
<td>ALA</td>
<td>15.9989996</td>
<td>O</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>NME</td>
<td>14.0107</td>
<td>N</td>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>NME</td>
<td>1.00800002</td>
<td>H</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CH3</td>
<td>NME</td>
<td>12.0109997</td>
<td>C</td>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HY13</td>
<td>NME</td>
<td>1.00800002</td>
<td>H</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HY12</td>
<td>NME</td>
<td>1.00800002</td>
<td>H</td>
<td>21</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HY13</td>
<td>NME</td>
<td>1.00800002</td>
<td>H</td>
<td>22</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Legend:
- Red: atom pair includes at least one hydrogen atom.
- Blue: heavy atom pair with less than two heavy atoms between them.
- Dark blue: heavy atom pair with exactly two heavy atoms between them.
- Light blue: heavy atom pair with more than two heavy atoms between them.

Classification of atom pairs for N-acetyl-N′-methylalaninamide by separation on chain and atom type (heavy atom or hydrogen)
Table 4.3  Summary of convergence for C7ax well → C7eq well BVP’s

| mode selection | H₀ (upper bound) | H on initial iteration | || | M¹²F¹² on initial iteration | || | M¹²F¹² on initial iteration | || | M¹²F¹² on initial iteration |
|----------------|------------------|------------------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| all            | 40 0 -10         | 40.0 15.0 40.0         | 5.4           | 4.4           | 3.0           | 4.0           | 1.0           | 0.9           | 1.0           | 0.9           |
| slowest 30     | 40 0 -10         | 57.4 23.2 41.1         | 3.6           | 2.7           | 3.7           | 4.0           | 1.0           | 0.9           | 1.0           | 0.9           |
| slowest 20     | 40 0 -10         | 53.5 15.8 46.8         | 5.2           | 3.3           | 2.2           | 4.0           | 1.0           | 0.9           | 1.0           | 0.9           |
| slowest 10     | 40 0 -10         | 14.3 4.3 24.0          | 4.2           | 0.5           | 2.4           | 4.0           | 1.0           | 0.9           | 1.0           | 0.9           |
| slowest 9      | 40 0 -10         | 14.4 4.0 24.4          | 4.1           | 0.4           | 2.4           | 4.0           | 1.0           | 0.9           | 1.0           | 0.9           |
| slowest 8      | 40 0 -10         | 14.0 3.9 24.0          | 4.4           | 0.4           | 2.4           | 4.0           | 1.0           | 0.9           | 1.0           | 0.9           |
| slowest 7      | 40 0 -10         | 13.6 3.3 22.0          | 4.4           | 0.4           | 2.8           | 4.0           | 1.0           | 0.9           | 1.0           | 0.9           |
| slowest 6      | 40 0 -10         | 11.3 3.4 20.4          | 1.8           | 0.4           | 1.3           | 4.0           | 1.0           | 0.9           | 1.0           | 0.9           |
| slowest 5      | 40 0 -10         | 5.6 2.9 19.8           | 1.9           | 0.9           | 1.4           | 4.0           | 1.0           | 0.9           | 1.0           | 0.9           |

Summary of convergence results for C7ax well → C7eq well BVP’s
Figure 4.1 $\phi$-$\psi$ contour plot: $C_{7_{eq}}$ potential energy well of $N$-acetyl-$N'$-methylalaninamide

$\phi$-$\psi$ contour plot: $C_{7_{eq}}$ potential energy well for $N$-acetyl-$N'$-methylalaninamide

The region enclosed by the thick black line defines combinations of values for the $\phi$ and $\psi$ angles which correspond to the $C_{7_{eq}}$ potential energy well.
Figure 4.2  $\phi$-$\psi$ plots for low total energy simulations
(a) $C_{7eq}$ potential energy well
Duration: 0.6 $ps$

$\phi$-$\psi$ plots for constant low energy simulation at different energy levels (in kcal mol$^{-1}$) beginning in $C_{7eq}$ for 0.6 $ps$
Figure 4.2 $\phi$-$\psi$ plots for low total energy simulations

(b) C7$_{za}$ potential energy well

Duration: 0.6 ps

$\phi$-$\psi$ plots for constant low energy simulation at different energy levels (in kcal mol$^{-1}$) beginning in C7$_{za}$ for 0.6 ps
Figure 4.3   Interatomic distance ranges by conformation type

Interatomic distance ranges by conformation type

Interatomic distance ranges by conformation type

\[ \text{atom pair index} \]

\[ \text{distance in Angstroms} \]

\[ \text{C7}_{\text{ax}} \quad \text{C7}_{\text{eq}} \]
Figure 4.4 $\phi$-$\psi$ plot of final trajectories
(a) all modes, slowest 30 modes, slowest 20 modes

$\phi$-$\psi$ plot of final trajectories: all modes, slowest 30 modes, slowest 20 modes
Figure 4.4 Φ-Ψ plot of final trajectories

(b) slowest 10 modes, slowest 9 modes, slowest 8 modes
Figure 4.4 $\phi$-$\psi$ plot of final trajectories

(c) slowest 7 modes, slowest 6 modes, slowest 5 modes

$\phi$-$\psi$ plot of final trajectories: slowest 7 modes, slowest 6 modes, slowest 5 modes
Figure 4.5 $\phi$-$\psi$ plot for initial trajectories: all modes

$\phi$-$\psi$ plot for initial trajectories: all modes
Figure 4.6 \( \phi-\psi \) plot for solution trajectories: all modes

\[ \phi-\psi \] plot for solution trajectories: all modes
Figure 4.7 $\phi$-$\psi$ plot for initial trajectories: slowest 5 modes

$\phi$-$\psi$ plot for initial trajectories: slowest 5 modes
Figure 4.8 $\phi$-$\psi$ plot for solution trajectories: slowest 5 modes
Figure 4.9  Distribution of total energy in $kcal \text{ mol}^{-1}$

Distribution of total energy in $kcal \text{ mol}^{-1}$ for final trajectories BVP problem set
Figure 4.10  Vibrational modes of C7ₐₓ local minimum in cm⁻¹

Vibrational modes of C7ₐₓ local minimum in cm⁻¹
5 DISTANCE MATRIX INTERPOLATION METHODS FOR BOUNDARY VALUE APPROACHES TO BIOMOLECULAR DYNAMICS SIMULATION

5.1 Abstract

Boundary value approaches to all-atom molecular dynamics simulation (AA-MDS) have been used to study the conformational transitions of proteins and other biomolecules. These approaches generally involve a mathematical formulation of a boundary value problem (BVP) for ordinary differential equations (ODE’s) and application of an iterative numerical method to solve the problem. In the realm of biologically interesting conformational transitions of proteins, boundary value approaches to AA-MDS must be selectively applied due to computational limitations. The range of values from which initial parameters for an iterative numerical method must be selected in order to achieve eventual convergence is limited, so the selection of appropriate initial parameters represents another challenge. The likelihood of convergence can be impacted by the method used for generating initial parameters. All-atom distance matrix interpolation (AA-DMI) methods for generating position trajectories that satisfy certain types of boundary conditions are less computationally demanding than boundary value approaches to AA-MDS, but do provide atomically detailed trajectories. These methods involve an optimization problem with an objective function derived by interpolation of interatomic distances between their values in one conformation and their values in another conformation. They can be expected to generate position trajectories that satisfy specified boundary conditions, but do not necessarily satisfy Newton’s equations of motion. We consider conformational transitions in the alanine dipeptide, \(N\text{-acetyl-N'}\text{-methylalaninamide}\), and identify some of the difficulties with an all-atom version of a DMI method that was previously introduced as an elastic network model (ENM). We introduce another AA-DMI method based on ideas and methods commonly used in molecular distance geometry (DG) and multidimensional scaling. We also
propose the use of interpolation by spline functions as an alternative to more the conventional and easily obtained interpolation by a linear polynomial. Refinement of AA-DMI position trajectories by constrained energy minimization is also proposed. Results are presented from the study of conformational transitions of an alanine dipeptide.

5.2 Introduction

All-atom molecular dynamics simulation (AA-MDS) is an established means to study molecular dynamics as it is widely accepted that if appropriate modeling assumptions are employed, molecular motions obtained from AA-MDS are useful depictions of molecular motions. AA-MDS has been used to study the conformational transitions of proteins and other biomolecules. But, AA-MDS is impacted by a system size limitation and a time interval limitation, both described in section 2.3.9 of this dissertation. These limitations stifle the direct application of AA-MDS to the study of some molecular conformational transitions of biological interest and motivate the development of alternative, approximate methods that are less computationally demanding than AA-MDS. In comparison with AA-MDS, approximate methods are generally not as closely linked to an accepted theory of molecular motion. Different approximate methods have different features. In contrast with AA-MDS, some approximate methods may not produce trajectories at the atomic level of detail, some may only provide a trajectory on a coarse mesh, not the fine mesh that is required for an AA-MDS trajectory, some may provide the position trajectory but not the velocity trajectory. Some methods that only provide a position trajectory also might not include the dynamic component meaning that an ordered set of conformations is obtained, but neither the length of the time interval, nor the mesh, or set of points in time at which each particular conformation is realized, is known. We call approximate position trajectories that do not include the dynamic component hidden time trajectories. If an approximate method provides only the position trajectory $x^{prx}(t)$, but the dynamic component is known, the velocity trajectory can be approximated using the fact that the velocity trajectory, $v^{prx}(t)$, satisfies $v^{prx}(t)=\frac{d}{dt}(x^{prx}(t))$ and, therefore, can be approximated by a finite difference approximation of the derivative. The accuracy of one of these finite difference approximations is
dependent on the local density of the mesh. If an approximate method generates a hidden time position trajectory, the method does not produce a velocity trajectory. Still, a sequence of velocity directions can be computed directly from a hidden time position trajectory and this sequence can be of practical use (see subsubsections 3.4.3.7 and 3.4.3.8).

As was described in subsection 2.4.2, the boundary value approach to AA-MDS, or boundary value AA-MDS (BV-AA-MDS), has some appeal as an approach for the study of conformational transitions. BV-AA-MDS, in general, requires an iterative numerical method for solving boundary value problems (BVP’s) for ordinary differential equations (ODE’s). To begin, these methods require an initial trajectory that might not satisfy the equations of motion or the boundary conditions of the BVP. The initial trajectory (and also trajectories of later iterations) are generated by a numerical algorithm that takes a set of parameters as input. A challenge in successfully applying BV-AA-MDS is that there may be a limited range of values for initial parameters that result in eventual convergence. So, convergence to a solution trajectory can be strongly dependent on the initial parameters. Relatively inexpensive alternative simulation or modeling methods may also be useful in complementing BV-AA-MDS methods when they are applied to generate initial trajectories for BV-AA-MDS. The likelihood of convergence can be impacted by the method used for generating initial parameters. Some specific uses of approximate methods for the purpose of generating initial trajectories for the multiple shooting method are described in portions of subsection 3.4.3.

In this chapter, we focus on a particular class of approximate methods which we collectively label as all-atom distance matrix interpolation (AA-DMI) method. This approximation method produces a hidden time position trajectory with atomic detail. It does not produce a velocity trajectory. For one type of AA-DMI method, the mesh density does not affect results in the sense that determination of each intermediate conformation is determined independently; for another type of AA-DMI method, a recursive method, the mesh density does affect results and it must be sufficiently dense in order to achieve reasonable results. Section 2.3.10 of this dissertation contains a short introduction and examples intended to provide the reader with an intuitive understanding of DMI methods. Because a DMI trajectory does not produce a
velocity trajectory, when we refer to a DMI trajectory in this chapter, it is implied that it is only a position trajectory.

In [Kim2002a] and [Kim2002b], a coarse-grained DMI method was introduced. It was described as a method that incorporates an elastic network model (ENM). The method was applied at a coarse-grained level of resolution meaning that a group of atoms (e.g. those within a single residue in a protein) were represented by a single point mass such as the C$^\alpha$ atom of a protein. In [Kim2003], trajectories generated using the DMI-ENM method were compared to initial-value molecular dynamics trajectories and were judged to be similar. The method can be applied at the all-atom level by letting a point correspond to an atom rather than a residue.

In DMI methods, a set of atom pairs are selected. For each selected atom pair, a target distance is set for a specified number of intermediate conformations that will form a trajectory. The target distances are generated by an interpolation method that takes into account the actual distances in the beginning conformations and ending conformations. In considering different interpolation methods, information about some intermediate conformations of the trajectory can be useful.

Before moving on to the main focus of this chapter — the generation of approximate trajectories by interpolation of distance matrices — we address a question that the reader might be asking: Why use a method based on interpolation of distance matrices rather than a method based on the simpler and perhaps more intuitive interpolation of rectangular coordinates? These two forms of interpolation have been contrasted in [Kim2002a]. It was concluded that linear interpolation of rectangular coordinates can result in trajectories with bond lengths and angles of the intermediate conformations that are unrealistic and permit atoms to occupy the same location in space at the same time. On the other hand, in DMI, relative distance between atoms are interpolated, so unrealistic conformations and superimposition of atoms are less likely. For further comparison with alternate methods of interpolation, see [Kim2002a].
5.3 Preliminaries: All-atom distance matrix interpolation using an elastic network model with linear polynomial interpolation for target distance matrices (linear AA-DMI-ENM)

In the ENM approach to DMI, the intermediate conformations that comprise a DMI position trajectory are generated incrementally using a recursive formula. Let \([t_0, t_f]\) be an interval of time over which a conformational transition occurs where we can assume \(t_0\) is zero, but \(t_f\) is unknown. Also, let \(\Delta t\) be a mesh, also unknown, such that such that \(\Delta t = \{t_k: 0 \leq k \leq D\}\) and \(0 = t_0 < t_1 < \ldots < t_D = t_f\). And, let \(x^{ENM}(\Delta t)\) be a \(3n \times (D+1)\) position trajectory on \(\Delta t\) determined by AA-DMI-ENM. The intermediate conformation \(x^{ENM}(t_{k+1})\) is determined by

\[
x^{ENM}(t_{k+1}) = x^{ENM}(t_k) + \dot{x}^k, \quad 0 \leq k \leq D
\]

where \(\dot{x}^k\) is the increment from the \(k^{th}\) conformation to the \(k+1^{th}\) conformation. In subsection 5.3.2 below, we describe the ENM approach for computing \(\dot{x}^k\). When the context is clear, we will drop the ENM superscript for notational simplicity.

5.3.1 Defining an optimization problem for linear distance matrix interpolation

Linear DMI is DMI in which the evolution of all the interatomic distances proceeds monotonically with uniform relative rate. To begin the description of the linear DMI approach for computing \(\dot{x}^k\), let \(w\) and \(z\) represent the \(3n \times l\) position vectors of known rectangular coordinates of the beginning and ending conformations for a conformation transition. Notationally, we write

\[
w = [w_{11} \mid w_{12} \mid w_{13} \mid w_{21} \mid w_{22} \mid w_{23} \mid \ldots \mid w_{i1} \mid w_{i2} \mid w_{i3} \mid \ldots \mid w_{n1} \mid w_{n2} \mid w_{n3}], \quad 1 \leq i \leq n
\]

with the second number of the subscripts corresponding to the coordinate directions. Let \(\tilde{w}\) represent an \(n \times 3\) representation of the rectangular coordinates. Using the notation \(w_i = [w_i1, w_{i2}, w_{i3}]\) to represent a row vector containing the rectangular coordinates of the \(i^{th}\) atom, we may write

\[
\tilde{w} = [w_{i1} \mid w_{i2} \mid \ldots \mid w_{i1} \mid \ldots \mid w_{i3}], \quad 1 \leq i \leq n
\]
Analogous representations also apply to \( z \).

Selected atom pairs are connected by springs with elasticities that we specify by a set of interpolation weights, or spring constants. The spring constant for the pair of atoms indexed by \( i \) and \( j \) is labeled \( k_{ij} \). The spring constant are arranged as the non-zero entries of a symmetric non-negative \( n \times n \) matrix \( \mathbf{K} \). Given a conformation, \( x \), in rectangular coordinates, a distance matrix, \( D(x) \), is an \( n \times n \) matrix in which the \( ij \)th entry, \( d_{ij}(x) \) contains the distance between atoms \( i \) and \( j \). The Euclidean metric will be assumed here, so

\[
(5.4) \quad d_{ij}(x) = \| x_i - x_j \|_2
\]

Let \( D(\Delta t) \) represent a sequence of distance matrices for the trajectory \( x^{ENM}(\Delta t) \). The idea of linear DMI is to apply entrywise linear interpolation of \( D(w) \) and \( D(z) \) to generate a set of target distance matrices on \( \Delta t \) that we will label as \( L(\Delta t) \). It is worthwhile to note that for any \( t_k \) where \( k \) is a nonnegative integer such that \( 0 < k < \Delta t \), \( L(t_k) \) may be an inconsistent distance matrix meaning that the set of distances may contain subsets which violate the triangle inequality. The \( ij \)th entry of \( L(t_k) \) is given by the formula.

\[
(5.5) \quad l_{ij}(t_k) = (1 - k/n) \| w_i - w_j \|_2 + (k/n) \| z_i - z_j \|_2
\]

There is an optimization problem to solve to determine each intermediate conformation (i.e. there is an optimization problem for each \( k \)). The objective function is

\[
(5.6) \quad C(D(x(t_k)), L, N) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=i+1}^{n} k_{ij} (d_{ij}(x(t_k)) - l_{ij}(t_k))^2
\]

and the optimization problem is

\[
(5.7) \quad \min_{x(t_k) \in \mathbb{R}^n} C(D(x(t_k)), L, N)
\]

The objective function \( C \) will also be referred to using the label \( \text{STRESS} \) due an equivalence with an objective function used in the field of multidimensional scaling (e.g.\cite{Tro1993},\cite{Hav1998}). Global minimization of \( \text{STRESS} \) is a nonlinear least squares optimization problem that, in general, is quite difficult to solve.
5.3.2 Simplifying the optimization problem by a quadratic approximation to the objective function

In what we call the ENM approach to optimization problem (5.7), we assume \( x(t_{k-1}) \) is known and think of \( x(t_k) \) as a small deviation \( \xi_k \) from \( x(t_{k-1}) \). So, \( x(t_k) = x(t_{k-1}) + \xi_k \). This implies that

\[
d_{ij}(x(t_k)) = ||x_i(t_k) - x_j(t_k)||^2 = ||x_i(t_{k-1}) + \xi_i(t_k) - x_j(t_{k-1}) - \xi_j(t_k)||^2.
\]

So, the objective function,

\[
C(D(x(t_k)), L, N),
\]

(5.8)

\[
C(\xi^k, x(t_{k-1}), L, N) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} ||x_i(t_{k-1}) + \xi_i(t_k) - x_j(t_{k-1}) - \xi_j(t_k)||^2
\]

and the optimization problem becomes

\[
(5.9) \quad \min_{\xi^k \in \mathbb{R}^{3n}} C(\xi^k, x(t_{k-1}), L, N)
\]

Since \( x(0) \) is known, this recursive approach has a starting point. Under the assumption that \( ||\xi^k||_2 \) is small, we can reasonably consider a quadratic approximation of \( C \) which we denote as \( \hat{C} \). It turns out that \( \hat{C} \) can be written in the form

\[
(5.10) \quad \hat{C}(\xi^k, x(t_{k-1}), L, N) = \frac{1}{2} (\xi^k)^T \Gamma \xi^k + \frac{1}{2} \gamma^T \xi^k + b
\]

where \( \Gamma \) is a \( 3n \times 3n \) matrix, \( \gamma \) is a \( 3n \times 1 \) vector and \( b \) is a constant. Formulas for \( \Gamma \) and \( \gamma \) are given in [Kim2002a] in the case of uniform weights. Formulas for non-uniform weights are almost identical. The value \( \xi \) that minimizes \( \hat{C}(\xi^k, x(t_{k-1}), L, N) \) can be determined by the value of \( \xi \) where \( \xi \) is a solution of

\[
(5.11) \quad \Gamma \xi + \gamma = 0
\]

There are infinitely many solutions to this equation since \( \Gamma \) will have three zero eigenvalues. Unique solutions can be obtained by specifying three additional constraints. Examples, of additional constraints include (1) specifying the translation for one atom from \( t = t_{k-1} \) to \( t = t_k \) and (2) specifying the linear momentum of the system from \( t = t_{k-1} \) to \( t = t_k \).

The observation that \( \Gamma \) has three zero eigenvalues suggests that \( \hat{C} \) retains properties of \( C \) as is explained below. A unique solution to (5.14) can be obtained by specifying that the linear momentum of the molecule is zero from \( t = t_k \) to \( t = t_k + 1 \). That solution can be translated to another center of mass in any
direction and by any magnitude and the value of $C$ will remain unchanged. There is one degree of freedom for choosing the center of mass. For any center of mass the molecule can be rotated arbitrarily around its center of mass. But, in rotating around the center of mass, the distance from each atom to the center of the molecule does not change. Any rotation about the center of mass can be obtained as the net result of two rotations in specified directions. (This can be clarified by focusing on one atom and considering possible locations of the atom in terms of spherical coordinates with an origin at the center of mass and with the radius coordinate fixed.) So, there are two degrees of freedom in choosing an arbitrary rotation. In total there are three degrees of freedom that need to be specified to determine a unique solution.

5.3.3 Generation of AA-DMI-ENM trajectories in three-dimensional space

For $w$ and $z$ representing the $3n \times 1$ position vectors of known rectangular coordinates of the beginning and ending conformations for a conformation transition, assume that an AA-DMI-ENM trajectory has rectangular coordinates at $x(t_0) = w$ that satisfy beginning conditions and internal coordinates of $x(t_f)$ are approximately equal to internal coordinates of $z$. Incremental translation and rotation may be required to give a smooth transition spatially and to satisfy $x(t_f) \approx z$. Below is a method for incremental translation and rotation.

First, generate an AA-DMI-ENM trajectory, $x^{ENM}(\Delta t)$, as described in the previous section.

Second, compute the center of mass of $w$, $\bar{w}^{(m)}$, and the center of mass of $z$, $\bar{z}^{(m)}$. Then, the incremental linear momentum from step $t_k$ to $t_{k+1}$ is equal to $\frac{1}{\Delta t} \left( \bar{w}^{(m)} - \bar{z}^{(m)} \right)$. Third, for $0 \leq k \leq \Delta t$, perform

$$x^{ENM}(t_k) \mapsto ALIGN(m, w, x^{ENM}(t_k)) + \frac{k}{\Delta t} \left( \bar{w}^{(m)} - \bar{z}^{(m)} \right)$$

to generate a trajectory $x^{ENM}(\Delta t)$ has the desired linear momentum and is optimally aligned with the beginning conformation. To ensure appropriate rotation of the molecule, we will create a sequence of rotation matrices, $Q(\Delta t)$ and auxiliary vectors of rotation angles, $\theta_1(\Delta t)$, $\theta_2(\Delta t)$, and $\theta_3(\Delta t)$ that will permit a uniform incremental rotation of the overall rotation between $w$ and $z$. To begin, let

$$Q(t_f) = R_{min}(m, z, w) = \begin{bmatrix} q_{11} & q_{12} & q_{13} \\ q_{21} & q_{22} & q_{23} \\ q_{31} & q_{32} & q_{33} \end{bmatrix}$$
Note that $Q(t_f)$ is a rotation matrix which can be written as

$$Q(t_f; \theta_1(t_f), \theta_2(t_f), \theta_3(t_f)) = Q_1(\theta_1(t_f)) Q_2(\theta_2(t_f)) Q_3(\theta_3(t_f))$$

where, dropping the argument $t_f$ for the inner functions for simplicity,

$$Q_1(\theta_1) = \begin{bmatrix} \cos(\theta_1) & -\sin(\theta_1) & 0 \\ \sin(\theta_1) & \cos(\theta_1) & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

$$Q_2(\theta_2) = \begin{bmatrix} \cos(\theta_2) & 0 & -\sin(\theta_2) \\ 0 & 1 & 0 \\ \sin(\theta_2) & 0 & \cos(\theta_2) \end{bmatrix},$$

$$Q_3(\theta_3) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta_3) & -\sin(\theta_3) \\ 0 & \sin(\theta_3) & \cos(\theta_3) \end{bmatrix},$$

Then, $\theta_1(t_f), \theta_2(t_f), \text{ and } \theta_3(t_f)$ can be determined sequentially by the formulas:

$$\theta_2(t_f) = \arcsin(q_{31}),$$

$$\theta_1(t_f) = \arcsin(q_{31} / \cos(\theta_2)),$$

$$\theta_3(t_f) = \arcsin(q_{32} / \cos(\theta_2)).$$

Now, for $0 \leq k \leq N$, let

$$\theta_1(t_k) = \frac{k}{N} \theta_1(t_f); \theta_2(t_k) = \frac{k}{N} \theta_2(t_f); \theta_3(t_k) = \frac{k}{N} \theta_3(t_f)$$

and let $Q(t_k)$ be the rotation matrix determined by

$$Q(t_k) = Q_1(\theta_1(t_k)) Q_2(\theta_2(t_k)) Q_3(\theta_3(t_k))$$

Finally, perform

$$x^{ENM}(t_k) \mapsto x^{ENM}(t_k) Q(t_k)$$

to generate a trajectory $x^{ENM}(\Delta')$ that incrementally accomplishes the specified translation and rotation.

### 5.3.4 Strategies for selection of subset of interatomic distances

In the AA-DMI-ENM approach, there are different possibilities for selecting atom pairs to be included in the distance matrix interpolation objective function. One strategy is to select atom pairs in which the two atoms are within a pre-specified cutoff distance in at least one of the two boundary conformations. A similar strategy is to select atom pairs that are within a pre-specified cutoff distance in...
both of the two boundary conformations, but it is less effective in coarse-grained DMI-ENM according to [Kim2002a]. These strategies do not seem appropriate for the study of a small molecule where all atoms are relatively close to each other. For example, in an alanine dipeptide all atoms are relatively close to each other. Regarding the selection of atom pairs to include in DMI calculations, one possibility is to select all atom pairs. Another possibility is to select just atom pairs in which both atoms are heavy atoms. In preliminary experiments for an alanine dipeptide, the results did not seem to vary much. In the results presented in the next section (section 5.4) for that alanine dipeptio, all atom pairs were selected.

5.3.5 Strategies for assignment of interpolation weight, or spring constants

The simple approach of assigning uniform interpolation weights, or spring constants, was shown to be effective in [Kim2002a], [Kim2002b], and [Kim2003]. It is desirable that all the intermediate conformations of a trajectory generated with an approximate method like AA-DMI have a reasonably low potential energy based on evaluation using a potential energy function like AMBER ([Cor1995]). Sometimes, some intermediate conformations of an AA-DMI trajectory exhibit seemingly slight deviations from preferred values for bond lengths and bond angles so that trajectories that otherwise seem feasible sometimes have some intermediate conformations with high energy values. The deviations from preferred values for bond lengths and bond angles can be primarily attributed to the local arrangement of atoms with fewer than two atoms between them. In preliminary experiments for an alanine dipeptide, appropriate selection of the matrix of spring constants sometimes helped to alleviate this problem, specifically by assigning relatively large weights for atom pairs separated by less than two atoms. A selection of spring constants that consistently alleviated this problem was elusive. The preliminary work also included DMI with mass-weighted spring constants. So, the spring corresponding to the \(ij^{th}\) entry of \(K\) is the product of the square root of the mass of the \(i^{th}\) atom times the square root of the mass of the \(j^{th}\) atom. In the end, the results presented in the next section were limited to uniform spring constants as there was no clear improvement from this choice. Each spring constant was assigned a unit value. An approach for choosing
spring constants based on chemical bond information was proposed in [Jeo2006]. Incorporating this type of information could be considered for future work.

5.4 Ideas, methods and analysis

The twenty-two atom dipeptide \(N\)-acetyl-\(N'\)-methylalaninamide and the potential energy function that we have used in the study of this molecule were introduced earlier in this dissertation. In this chapter, we use this alanine dipeptide again as an example. The analysis again considers the C-N-C{\(\alpha\)}-C dihedral angle (\(\varphi\)) and the N-C{\(\alpha\)}-C-N dihedral angle (\(\psi\)) of the alanine residue as these angles are of primary importance in determining the conformation of the alanine dipeptide. Another dihedral angle of significance in transitions between C7{\(\alpha\)} and C7{\(\alpha\)} potential wells that was identified in [Bol2000] is O-C-N-C{\(\alpha\)} and it will be labeled as \(\theta\) here (see Figure 5.1). \(\theta\) should not be confused with the other type of backbone dihedral angle C{\(\alpha\)}-C-N-C{\(\alpha\)} which is commonly labeled as \(\omega\) (or with the angles \(\theta_1\), \(\theta_2\), and \(\theta_3\) of subsection 5.3.3). We will apply distance matrix interpolation methods to study conformational transitions from the primary local minimum of the C7{\(\alpha\)} potential energy well to the primary local minima of the C6, C5{\(\beta\)}, and C7{\(\alpha\)} potential energy wells.

5.4.1 Identifying and evaluating assumptions of linear DMI-ENM

In the previous section, a method for generating a linear DMI-ENM trajectory was described. Because the quadratic approximation of \(C\) is based on the assumption that \(z^k\) is a small deviation, this method will not give useful results if the mesh is too coarse. Upon termination of the recursive algorithm, the internal coordinates of \(x^{ENM}(t_f)\) should be approximately equal to the internal coordinates of \(z\). Equivalently, the relation \(RMSD(m; x^{ENM}(t_f), z) \approx 0\) should hold, but it should be verified since it cannot be guaranteed. What are the implications if this relation is not satisfied to a desired tolerance? The mesh could be too coarse. But, the validity of a couple of assumptions of the linear DMI-ENM method are also worth some scrutiny. First, the set of target distance matrices, \(L(\Delta')\), is generated based on the assumption that optimal trajectories are those for which interatomic distances for each selected pair progress at about the same rate relative to the total change in distance between the beginning and ending conformation for the
selected pair. We will call this assumption the *linearity assumption*. From the perspective of the author, it does not seem that the linearity assumption would be appropriate for all transition pathways. Second, the assumption that $C$ can be adequately approximated by $\hat{C}$ will be called the *quadratic approximation assumption*. This assumption might not always be appropriate.

It would be useful to isolate and evaluate the legitimacy of these assumptions separately. For the linearity assumption, consider a BVP with boundary conditions that require a conformational transition as a solution and suppose that a BV-AA-MDS solution or approximate BV-AA-MDS solution exists. To qualitatively assess the uniformity of relative change in interatomic distances for a given conformation, for any atom pair, one can view a plot of the interatomic distance versus time. Plots of this type will be called *dynamic distance distribution (DDD) plots*. If the linearity assumption is satisfied, then the graphs of the DDD plots should all be monotone with similar (though possibly inverted) patterns of curvature. Examining the DDD plots for these qualities can help in assessing the validity of the linearity assumption.

With respect to the quadratic approximation assumption, consider a BVP with boundary conditions that require a conformation transition as a solution and suppose that an BV-AA-MDS solution exists. Apply a modified version of the linear AA-DMI method in which the set of target distance matrices, $L(\Lambda')$, is replaced by the set of actual distance matrices, $D^{MDS}(\Lambda')$ from the BV-AA-MDS solution trajectory. This modified version is effectively a specific type of nonlinear AA-DMI method. If the quadratic assumption is valid, it would seem reasonable to expect that the resulting AA-DMI trajectories would be similar to or, ideally, identical to the BV-AA-MDS solution position trajectories. In this way, the validity of the quadratic approximation assumption can be investigated.

### 5.4.2 Application of linear AA-DMI-ENM to conformational transitions of *N*-acetyl-*N'*-methylalaninamide

We consider linear DMI for each of the three transitions — $C_{7\text{eq}} \rightarrow C_6$, $C_{7\text{eq}} \rightarrow C_{5\beta}$, and $C_{7\text{eq}} \rightarrow C_{7\text{ax}}$ — being studied. One useful way of analyzing the trajectories is visualization using visualization software like Visual Molecular Dynamics (VMD). Other useful means of analysis include time series plots of values
of the important dihedral angles ($\phi$, $\psi$, and, $\theta$), potential energy ($U(x_{ENM}(t_i))$, and distance matrix error from the ending conformation ($DME(x_{ENM}(t_k), x_{ENM}(t_f))$) for $0 \leq k \leq D$. If the linear ENM method is to provide feasible trajectories, at minimum, the dihedral angle time series plots should be reasonably smooth. Any feasible trajectory should not have excessively high values of potential energy. And the $DME$ values are expected to decrease smoothly and more or less monotonically to zero.

### 5.4.2.1 Feasibility of the linear AA-DMI-ENM trajectories

In Figure 5.2, the values of $\phi$ (black), $\psi$ (magenta), $\theta$ (dotted gray), $U(x_{ENM}(t_i))$ (dashed dotted red), and $DME(x_{ENM}(t_i), x_{ENM}(t_f))$ (dashed green) are superimposed in overlay plots for the three transitions. We will call these $\phi/\psi/\theta/DME/U$ plots. Subplots (a), (b), and (c) correspond to the three transitions — $C7_{eq} \rightarrow C6$, $C7_{eq} \rightarrow C5_{\beta}$, and $C7_{eq} \rightarrow C7_{ax}$ — respectively. Figure 5.2a and Figure 5.2b suggest that the AA-DMI-ENM trajectories for $C7_{eq} \rightarrow C6$, $C7_{eq} \rightarrow C5_{\beta}$ represent feasible conformational transitions. On the other hand, the quasi-discontinuous ‘jumps’ in the various line plots of Figure 5.2c, however, suggest that the AA-DMI-ENM trajectory for $C7_{eq} \rightarrow C7_{ax}$ is not feasible and the evolution of the trajectory contains unrealistic, non-physical movements. For comparison purposes, in Figure 5.3a-c, for each of the three transitions, the same overlay plots are shown for BV-AA-MDS solution trajectories.

### 5.4.2.2 Assessing the linearity assumption

It is reasonable to ask whether or not the infeasibility of a linear ENM trajectory is correlated with violations of the linearity assumption. DDD plots for atom pairs in which both atoms are heavy atoms and the two atoms are separated on the chain by at least two other molecules are used to assess the linearity assumption. Figure 5.4, Figure 5.5, and Figure 5.6 show DDD plots for the $C7_{eq} \rightarrow C6$, $C7_{eq} \rightarrow C5_{\beta}$, and $C7_{eq} \rightarrow C7_{ax}$ transitions, respectively. Two additional DDD plots are shown for the $C7_{eq} \rightarrow C7_{ax}$ transition—a second approximation solution (Figure 5.7) and an exact solution (Figure 5.8). These additional plots are shown to give an idea of the variation in dynamic distance distributions for similar trajectories. Comparing DDD plots for approximate and exact solutions in Figure 5.7 and Figure 5.8, there are minor, but
unanticipated changes in the shapes of some of the DDD curves. For example, consider the DDD plots for the CH3-ACE:C-NME pair, the O-ACE:C-NME pair, and the C-ACE:C-NME pair.

Visual inspection of the DDD plots for these trajectories suggest that, qualitatively speaking, the approximate $C7_{eq} \rightarrow C6$ trajectory exhibits relative linearity in terms of progression of heavy atom interatomic distances with respect to time while the $C7_{eq} \rightarrow C5_{\beta}$ trajectory and the $C7_{eq} \rightarrow C7_{ax}$ trajectories exhibit relative nonlinearity. There is some evidence of correlation between the linearity exhibited in these plots and results for linear DMI-ENM as the linear DMI-ENM trajectory of the plot of Figure 5.2a is not problematic for the $C7_{eq} \rightarrow C6$ transition while the linear DMI-ENM trajectory of Figure 5.2c for the $C7_{eq} \rightarrow C7_{ax}$ transition is. We must note that the linear DMI-ENM trajectory of Figure 5.2b for the $C7_{eq} \rightarrow C5_{\beta}$ trajectory is not problematic, either. Informal analyses of a collection of linear ENM trajectories generated using various different spring constant matrices provides that the likelihood of a problematic trajectory, in order from lowest to highest by transition is lowest for $C7_{eq} \rightarrow C6$ followed by $C7_{eq} \rightarrow C5_{\beta}$ and then $C7_{eq} \rightarrow C7_{ax}$ (data not shown). Particularly if this assertion is valid, there is some evidence that there is a correlation between nonlinearity in progression of heavy atom interatomic distances and difficulties in linearly interpolated AA-DMI-ENM trajectories. Further analyses would be useful to provide a clearer picture of this suggested relationship. To further illustrate the problem with the linear DMI-ENM model, compare the DDD plots of Figure 5.6, Figure 5.7, and Figure 5.8 for three $C7_{eq} \rightarrow C7_{ax}$ AA-MDS trajectories with the DDD plot of the problematic DMI-ENM trajectory of Figure 5.2c for the $C7_{eq} \rightarrow C7_{ax}$ transition. The DDD plot of this DMI-ENM trajectory is shown in Figure 5.9 assuming a uniformly distributed mesh.

5.4.2.3 **Assessing the quadratic approximation assumption**

We now assess the validity of the quadratic approximation to the objective function. For each of the three conformational transitions being studied, we applied nonlinear AA-DMI-ENM to four different conformational transitions from $C7_{ax}$ potential energy well to the desired ending potential energy well — C6, $C5_{\beta}$ or $C7_{ax}$. Each of these $3 \times 4 = 12$ sets of conformational transitions were defined based on AA-MDS
trajectories. The beginning and ending conformations of the AA-MDS trajectories were selected as beginning and ending conformations for application of the nonlinear AA-DMI-ENM method. The target distance matrices were defined as the distance matrices of the actual AA-MDS trajectories.

In Figure 5.10, overlay plots are shown for the twelve DMI-ENM trajectories generated with the sequence of target distance matrices equal to distance matrices of the actual MDS trajectories. This figure provides evidence that all four of the $C7_{eq} \rightarrow C6$ DMI-ENM trajectories were feasible, but one of the four $C7_{eq} \rightarrow C5_{\beta}$ DMI-ENM trajectories was problematic and all four of the $C7_{eq} \rightarrow C7_{ax}$ DMI-ENM trajectories were problematic. Since all of the target distance matrices were consistent, all twelve of the DMI-ENM trajectories should have been feasible. The fact that five of the twelve were not suggests that the quadratic approximation to the objective function is problematic.

### 5.4.2.4 Additional considerations

In further attempts to resolve these difficulties, different weighting strategies and different atom selections strategies were employed (results not shown). Performance can depend on the choice of weight matrix, but not in what seems to be a predictable manner. So, our experience has been that resolving these unrealistic movements by modification of the weight matrix, $\mathbf{X}$, is not a successful remedy. We also attempted to identify and describe the problematic movements by characterizing some tendencies of the problematic AA-DMI-ENM trajectories. This led to a revised algorithm which attempted to address the problems by re-initializing the AA-DMI-ENM algorithm in the iteration preceding a problematic iteration. In other words, a ‘new’ AA-DMI-ENM trajectory was created beginning near the problematic part of the trajectory and the ending conformation unchanged. Upon creation of this trajectory, it was spliced on the end of the partially completed original trajectory. This algorithm sometimes gave feasible results, but in general, the results were inconsistent.
5.4.3 An alternative approach to DMI: All-atom distance matrix interpolation using the STRAIN() objective function

It seems that the quadratic approximation, \( \hat{C} \), to \( C \) is inadequate. Let us consider another approach to DMI that will not require the quadratic approximation assumption. If a target distance matrix is consistent, this method will generate a feasible set of atomic coordinates that satisfies all of the target distances. And, if a target distance matrix is inconsistent, the set of atomic coordinates that is generated for that distance matrix will, in a measurable sense, optimally fit that inconsistent distance matrix. Additionally, this method will not involve a recursive approach thereby avoiding potential problems due to accumulation of errors from a series of quadratic approximations.

In the process of describing the ENM approach to DMI, the minimization problem (5.7) with the objective function,

\[
(5.24) \quad \text{STRESS}(D(x), L, \aleph) = \frac{1}{2} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} N_{ij}^2 \left( d_{ij}^2(x) - l_{ij}^2 \right)
\]

arose. This minimization problem needed to be solved for each intermediation conformation, \( x \), of a DMI trajectory. Even though this objective function seems appropriate, the problems associated with the quadratic approximation assumption can be eliminated indirectly by modifying the objective function. Two other objective functions will be introduced in this section. First, there is the SSTRESS() objective function. It takes the form

\[
(5.25) \quad \text{SSTRESS}(D(x), L, \aleph) = \frac{1}{2} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} N_{ij}^2 \left( d_{ij}^2(x) - l_{ij}^2 \right)^2
\]

SSTRESS() and STRESS() are similar functions. SSTRESS() is characterized as the smoother function; the leading \( S \) represents the worth ‘smoothed’. STRESS() is a summation of squared differences between squared actual distances and squared target distances while STRESS() is a summation of squared differences between actual distances and target distances. It is worth noting also that if \( L \) is a consistent distance matrix and \( x \) is a global minimizer of \( \text{STRESS}(D(x), L, \aleph) \) with \( \text{STRESS}(D(x), L, \aleph) = 0 \), then \( x \) is also a global minimizer of \( \text{SSTRESS}(D(x), L, \aleph) \) and \( \text{SSTRESS}(D(x), L, \aleph) = 0 \).
Now, consider the case of an inconsistent distance matrix for the minimization problem (5.7) defined with \( C = \text{SSTRESS}() \). Unfortunately, there are no known algorithms for finding efficiently finding global minimizers for \( \text{SSTRESS}() \) ([Tro1993],[Hav1998]); \( \text{SSTRESS}() \) is similar to \( \text{STRESS}() \) in this regard. But, we have not introduced \( \text{SSTRESS}() \) without reason. The reason is that there is another objective function, called \( \text{STRAIN}() \), that is closely related to \( \text{SSTRESS}() \) for which global minimizers can be reliably determined. It is instructive to examine the relationship between these two functions. Below is an adapted version of the presentation of this relationship in T. Havel’s “Distance geometry: theory, algorithms and chemical applications” found in *Encyclopedia of Computational Chemistry* ([Hav1998]). In order to do this, first rewrite \( \text{SSTRESS}() \) as

\[
\text{SSTRESS}(x, L, \mathcal{N}) = \frac{1}{2} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbf{S}_{ij} \left\| x_i - x_j \right\|^2 - l_{ij}^2 \]

Expanding \( \|x_i - x_j\|_2 \) and rearranging we can write this as

\[
\text{SSTRESS}(x, L, \mathcal{N}) = \frac{1}{2} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} 4\mathbf{S}_{ij} \left( x_i \cdot x_j - \frac{1}{2} \left( \|x_i\|^2 + \|x_j\|^2 - l_{ij}^2 \right) \right)^2
\]

where, in general, \( w \cdot z \) represents the dot product of two identically size vectors, \( w \) and \( z \). So, \( \text{SSTRESS}() \) can be written as a weighted sum of squares of the differences between dot products of the yet undetermined atomic coordinates in vector form and an estimate of these dot products that also includes undetermined quantities, namely, atomic distances from the origin, i.e. the set of distances between the individual atoms and the origin.

Now, let’s make four assumptions. First, assume the spring constant for each atom pair can be written as a product of two terms, one for each atom. So, write \( \mathbf{S}_{ij} = \mathbf{S}_i \mathbf{S}_j \). Second, assume the origin for the set of atomic coordinates to be determined will be the estimated center of mass of the system. Third, assume \( L \) is an estimate of \( D(x) \). Then, we can estimate atomic distances from the origin as follows. For \( 1 \leq i \leq n \), if \( l_{0i} \) represents the estimated distance of atom \( i \) from the origin and \( m \) represents the mass of the system, then \( l_{0i} \) is given by

\[
l_{0i}^2 = \frac{1}{\sum m_j} \sum_{j=1}^{n} m_j l_{ij}^2 - \frac{1}{\sum m_j} \sum_{j=1}^{k-1} \sum_{k=j+1}^{n} m_j m_k l_{jk}^2
\]
The fourth assumption is that \( l_0 \approx \|x_i\|_2 \).

The second objective function to be introduced in this section, called \( \text{STRAIN}(x, L, \mathbf{N}) \), incorporates the aforementioned assumptions. Define \( t_y = \frac{1}{2} t_{11} + \frac{1}{2} t_{22} - \frac{1}{2} t_{11} \). Then, the \( \text{STRAIN}(x, L, \mathbf{N}) \) objective function is

\[
(5.29) \quad \text{STRAIN}(x, L, \mathbf{N}) = \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{N}_{ij} (x_i \cdot x_j - t_y)^2
\]

While we have described how \( \text{STRAIN}(x, L, \mathbf{N}) \) can be derived from \( \text{SSTRESS}(x, L) \), how these two objective functions are related conceptually may not be obvious. To proceed toward a conceptual understanding, let \( \mathbf{N} \) be a diagonal matrix with the \( i \)th entry as \( N_{ii} \). And let \( L \) be a matrix with \( ij \)th entry \( l_{ij} \).

Then, \( \text{STRAIN}(x, L, \mathbf{N}) \) may be written in terms of the Frobenius norm of an \( n \times n \) matrix as

\[
(5.30) \quad \text{STRAIN}(x, L, \mathbf{N}) = \|L\mathbf{N} - \mathbf{X}^\top \mathbf{X}\|_F^2
\]

Furthermore, let \( \mathbf{1} \) be an \( n \)-component vector of ones, \( \mathbf{1} = [1; 1; \ldots; 1] \) and, in general, let \( w \ast z \) represent the Hadamard, or entrywise, product of \( w \) and \( z \). Then, the following relation holds:

\[
(5.31) \quad x^\top x - L \ast L = -\frac{1}{2} (1 - 1m^\top / \Sigma m)(D(x) \ast D(x) - L \ast L)(1 - 1m^\top / \Sigma m)
\]

Now, define \( P = (\sqrt{\frac{1}{2}})(1 - 1m^\top / \Sigma m) \). Then,

\[
(5.32) \quad \text{STRAIN}(x, L, \mathbf{N}) = \|N(P(D(x) \ast D(x) - L \ast L)P^\top \mathbf{N})\|_F^2
\]

Note that \( \text{STRESS}(x, L, \mathbf{N}) \) can be written, respectively as

\[
(5.33) \quad \text{STRESS}(x, L, \mathbf{N}) = \frac{1}{2} \|N(D(x) - L)\|_F^2
\]

and

\[
(5.34) \quad \text{SSTRESS}(D(x), L, \mathbf{N}) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=i+1}^{n} N_{ij} (d_{ij}^2(x) - l_{ij}^2)^2
\]

The \( n \times n \) matrix \( P \) in (5.32) functions as a two-sided projection matrix onto an \( n-1 \) dimensional subspace of \( \mathbb{R}^n \). Conceptually, because one degree of freedom was used to estimate the center of mass, the \( P \) appears in \( \text{STRAIN}(x, L, \mathbf{N}) \), but is otherwise identical to \( \text{SSTRESS}(x, L, \mathbf{N}) \).

A partial derivation and the formula for the global minimizer for \( \text{STRAIN}(x, L, \mathbf{N}) \) are given below. The details are not included here can be found in [Hav1998]. Let \( \mathbf{X} = N \mathbf{X} \) and let \( \mathbf{L} = N \mathbf{L} \). Then, \( \text{STRAIN}(x, L, \mathbf{N}) \) can be written as
Also, let $L = UΣU^T$ be the singular-value decomposition of $L$; let $A$ be a $3 \times 3$ diagonal matrix containing the three largest eigenvalues of $L$ — $λ_1$, $λ_2$, $λ_3$ — in the $1^{\text{st}}$, $2^{\text{nd}}$, and $3^{\text{rd}}$ diagonal entries; and let $V = [u_1; u_2; u_3]$ be an $n \times 3$ matrix where $u_1$, $u_2$, and $u_3$ are normalized independent eigenvectors for the eigenvalues $λ_1$, $λ_2$, and $λ_3$ respectively. Then, $\mathbf{x} = Α^{1/2}V^T\mathbf{x}$.

In review, for the purpose of determining one of the intermediate conformations of an AA-DMI trajectory, an optimization problem can be formulated. $STRESS()$ may be the most natural objective function, so its role as the objective function in the development of the DMI-ENM method in [Kim2002a] and [Kim2002b] is understandable. $SSTRESS()$ is also a reasonable choice of an objective function. Making some reasonable assumptions, $SSTRESS()$ can be replaced by $STRAIN()$. Let $η()$ be an operator that transform a set of coordinates from an $n \times 3$ representation to a $3n \times 1$ representation. Then, a trajectory generated using $C=STRAIN()$ in the minimization problem (5.7) on mesh, $Δ'$, is given by

$$x^{DG}(t_k) = η(Ν^{-1}VΔ^{1/2}) \quad \text{for} \ 0 ≤ k ≤ Ω$$

The DG superscript is meant represent ‘distance geometry’ as $STRAIN()$ have been widely used in applications of distance geometry. We will refer to a trajectory generated this way as a DG trajectory, or DMI-DG trajectory, or AA-DMI-DG trajectory.

### 5.4.4 Interpolation by cubic spline functions for target distance matrices

We now turn our attention to the linearity assumption. In section 5.3.1, an assumption about uniformity of relative change in interatomic distances was termed the linearity assumption. If this assumption is made, the set of target distance matrices, $L(Δ')$, is easy to generate. An intuitive way to relate the linearity assumption to the generation of $L(Δ')$ is to assume the components of $Δ'=[t_0; t_1; \ldots; t_{Ω-1}; t_Ω]$ are uniformly distributed on $Δ'$. Then, for each selected pair of atoms, $i$ and $j$, the target distance for any point of $Δ'$ is determined by a line that passes through $(t_0, d_{ij}(x(t_0)))$ and $(t_Ω, d_{ij}(x(t_Ω)))$. This amounts to interpolation of the interatomic distance for each selected atom pair by a linear polynomial. Note that even the linearity assumption could still be satisfied even if the components of $Δ'$ are not uniformly distributed.
If they are not, the set of target distances $L(\Delta')$ is still valid, but the linear graph described above would be replaced by some monotonic function. Assume that the linearity assumption does not always hold. It is reasonable to ask: Can a nonlinear method be used to produce a AA-DMI trajectory? Well, as long there is a way to produce a set of target distance matrices, $L(\Delta')$, on $\Delta'$, then methodologically nonlinear DMI is identical to linear DMI. The relevant question then with respect to nonlinear DMI is: What method will be used to generate $L(\Delta')$?

Here is one scenario in which a nonlinear method for generating $L(\Delta')$ seems reasonable. First, recall that in the simulation of a conformational transition using a two-point BVP, conditions are placed at the beginning and end of the time interval of the simulation based on knowledge about the beginning and ending conformations of the molecule. But, suppose in addition, knowledge about some intermediate conformations are available. Let the timepoints of the those intermediate conformations be given as the elements of a mesh, $\Delta' = [s_0; s_1; s_2; \ldots; s_N; s_{N+1}]$. (The values of the components of $\Delta'$ may possibly be unknown, but it is assumed the intermediate conformations can be placed in an ordered sequence.) So, let $x^{\text{MDT}}(\Delta')$ be a BV-AA-MDS solution trajectory with the known intermediate conformations. But, suppose the mesh for which a DMI trajectory is desired is more dense. For convenience, assume $\Delta' \subset \Delta'$.

One possible way to handle this is to apply linear DMI separately on the individual subintervals defined by $\Delta'$. This amounts to linear spline interpolation with $N + 1$ knots. Another possibility, though, is to use another interpolation method. In polynomial interpolation, for $N + 1$ support points, $(t_k, d_{ij}(x(t_k)))$, $k = 0, 1, 2, \ldots, N$, there exists a unique polynomial function $P$ such that and $P(t_k) = d_{ij}(x(t_k))$, $k = 0, 1, 2, \ldots, N$.

Linear DMI involves interpolation of the interatomic distance for each selected atom pair by a linear polynomial. So, linear DMI is a special case of polynomial interpolation with 2 support points (i.e. $N = 1$).

In the case of $N > 1$, a natural extension might be seem to perform interpolation of the interatomic distance for each selected atom pair by an $N^{th}$ degree polynomial. Interpolation by higher-order polynomials can sometimes yield curves with large oscillations that would not desirable for interatomic distances. Interpolation by cubic splines tend to yield smoother interpolating curves, and that is the interpolation method that will be employed. In particular, interpolation by cubic ‘not-a-knot’ splines will be employed.
The not-a-knot cubic spline function for interatomic distances of the atom pair with atoms indexed by \(i\) and \(j\) will be denoted by a real valued function \(S_{ij}\Delta s: [t_0,t_N] \to \mathbb{R}\) with the properties:

1. \(S_{ij}(s_k) = d_{ij}(x(s_k))\) for \(k=0,1,\ldots,N\)
2. \(S_{ij}\) is twice continuously differentiable on \([t_0,t_N]\)
3. \(S_{ij}\) coincides on every subinterval \([s_k,s_{k+1}], k=0,1,\ldots,N-1\), with a polynomial of degree at most three.
4. \(S_{ij}'''\) is continuous at \(t=s_j\) and \(t=s_{N-j}\). (This is the property that makes this cubic spline function a ‘not-a-knot’ cubic spline function.)

This cubic spline method is employed to form a smooth curve that gives the values of the interatomic distances on \(\Delta t\) for each atom pair based on an interpolation based on the values of the interatomic distances on \(\Delta s\). As a reminder, we seek an AA-DMI trajectory, \(x^{DMI}(\Delta t)\) that is intended to approximate a BV-AA-MDS solution trajectory. Assume that it approximates the particular solution trajectory \(x^{MDS}(\Delta t)\). For \(k=0,1,\ldots,N\), the cubic spline interpolation method will satisfy \(x^{DMI}(s_k) = x^{MDS}(s_k)\), so \(D(x^{DMI}(s_k)) = D(x^{MDS}(s_k))\). To determine \(x^{DMI}(\Delta t)\), we first must determine \(D(x^{DMI}(\Delta t))\). The target matrix \(L(\Delta t)\) will be determined by cubic spline interpolation. The sequence of \(ij^{th}\) entries of \(L(\Delta t)\) will be the values of the interpolating spline function of \(d_{ij}(x^{DMI}(\Delta s))\) on \(\Delta t\) with knots at the \(N+1\) components of \(\Delta t = [s_0; s_1; s_2; \ldots; s_N; s_{N+1}].\)

### 5.4.5 Refining trajectories using local constrained energy minimization (LCEM)

Even though an AA-DMI trajectory, \(x^{DMI}(\Delta t)\) might have some desirable features and may have been obtained by satisfying some requirements for optimality, it is still possible that some of the intermediate conformations may not be appropriate with respect to the potential energy function, in the sense that their energy levels may be high. Trajectories that have intermediate conformations of excessively high potential energy may not be desirable (e.g. if, for example, the AA-DMI trajectory is being used to generate low energy initial trajectories for a BV-AA-MDS algorithm). So, it may be
desirable to perform local constrained energy minimization using the conformations of \( x_{OM}^{\Delta} \) as starting points to find conformations that are sufficiently low in energy. It is important that the energy minimization for each conformation preserves essential properties of the initial conformation because the initial trajectory is an ordered sequence of conformations. One way to accomplish this is to perform the energy minimization with a set of constraints specified to preserves essential properties. In the next subsection, we provide an ad-hoc approach for the alanine dipeptide.

### 5.4.6 Application of spline interpolation and AA-DMI-DG to conformational transitions of \( N\text{-acetyl-N'}\text{-methylalaninamide} \)

We have hypothesized that problems resulting from two questionable assumptions of linear DMI-ENM can be addressed by considering DMI trajectories generated using the \( STRAIN() \) objective function and using target distance matrices determined by cubic spline interpolation. Additionally, we have hypothesized that refinement of some conformations of a trajectory by constrained energy minimization might be useful. In this section, we examine simulation results for the alanine dipeptide, \( N\text{-acetyl-N'}\text{-methylalaninamide in vacuo} \), that we used previously.

#### 5.4.6.1 Using the \( STRAIN() \) objective function with consistent target distance matrices

In subsubsection 5.4.2.3, evidence was provided that the quadratic approximation assumption of AA-DMI-ENM is problematic. Twelve AA-DMI-ENM trajectories were generated using a twelve sets of consistent target distance matrices. But, five of the twelve trajectories were not feasible. We now consider generating twelve AA-DMI-DG trajectories using the same twelve sets of consistent target distance matrices. Unlike the AA-DMI-ENM trajectories, the internal coordinates of the conformations of the AA-DMI-DG trajectories match the internal coordinates of the conformations of the MDS trajectory used to generate the target distance matrices. (see Figure 5.11). So, the problem of AA-DMI-ENM associated with the quadratic approximation assumption has been avoided.
5.4.6.2 Local constrained energy minimization (LCEM)

Below we describe a local constrained energy minimization approach to refine an AA-DMI trajectory, $x^{DMI}(\Delta t')$, for the alanine dipeptide:

1. Local constrained energy minimization (LCEM) is performed for each individual conformation $x^{DMI}(\Delta t')$. For each $k$, the conformation $x^{DMI}(t_k)$ was used as the starting conformation.

2. Unconstrained minimization is performed using the built-in MATLAB function, `fminunc()`. The target function is variation of the MATLAB version of the AMBER potential energy function. This variation includes a quadratic penalty term for selected dihedral angles. The stationary values for the quadratic penalty term were selected to be the values of the dihedral angles for the STRAIN minimized starting points.

3. Five dihedral angles were selected—including two improper dihedral angles.
   a. O-ACE-OOH : C-ACE-OOH : N-ALA : C$\alpha$-ALA (the $\theta$ angle)
   b. C-ACE-OOH : N-ALA : C$\alpha$-ALA : C-ALA-O (the $\phi$ angle for the alanine residue.)
   c. N-ALA : C$\alpha$-ALA : C-ALA-O : N-NH2 (the $\psi$ angle for the alanine residue).
   d. C$\alpha$-ACE C-ACE-OOH N-ALA C$\alpha$-ALA (an improper dihedral angle which indicates the non-planarity of the $\omega$ angle preceding the alanine residue)
   e. C-ALA-O N-NH2 O-ALA C$\alpha$-ALA (an improper dihedral angle which indicates the nonplanarity of the $\omega$ angle following the alanine residue)

Successful termination of the `fminunc()` function call provides one conformation of a refined trajectory that we denote by adding the -LCEM suffix to the trajectory name.

5.4.6.3 Spline interpolation for target distance matrices

Consider generating a series of four AA-DMI-DG trajectories along with a series of four corresponding AA-DMI-DG-LCEM trajectories for the $C_{7\text{eq}} \rightarrow C_{7\text{ax}}$ transition using a set of target distance matrices generated by cubic spline interpolation of interatomic distances with five knots. The five knots are the beginning and ending conformations and timepoints along with three intermediate conformations.
and timepoints taken from each of four BV-AA-MDS solution trajectories. In Figure 5.12, $\phi/\psi/\theta/DME/U$ plots are shown for each of the four AA-DMI-DG trajectories and, also, the four AA-DMI-DG-LCEM trajectories. These plots are similar to $\phi/\psi/\theta/DME/U$ plots for the BV-AA-MDS solution trajectories. In Figure 5.13, DDD overlay plots are shown corresponding to the 1\textsuperscript{st} of the four BV-AA-MDS solution trajectories. The DDD overlay plots include DDD plots of the BV-AA-MDS solution, the spline functions, the DG solution, and the DG-LCEM solution. These plots are shown with the same scale used on previous plots. On this scale, they are mostly indistinguishable. In Figure 5.14, the scales are changed appropriately to enable a closer look. At these higher resolutions, we are able to see some differences. First, observe that the spline curves and the spline-based DG curves are strongly correlated, suggesting that the set of distances determined by the cubic spline approach are nearly consistent in almost all cases. Second, we note that the solution trajectories exhibit some small quasi-periodic or anharmonic oscillations in interatomic distances that are not modeled by either DG or DG-LCEM. It would be useful to try to determine the results of spline interpolation using knots that are not directly taken from BV-AA-MDS solution trajectories. For example, the knots could contain conformations that are perturbations of conformations from actual solution trajectories or they could contain conformations generated using another approximate method.

### 5.4.7 Additional analysis and further discussion

Below are some additional comments regarding the study of the alanine dipeptide and its implications.

1. It was conjectured that $STRESS()$ might be the most natural objective function for DMI. Using $STRAIN()$ as the objective function instead of $STRESS()$ is convenient in that a global minimizer can be obtained without needing to resort to a quadratic approximation. But, in using $STRAIN()$ instead of $STRESS()$, the quadratic approximation assumption has been effectively replaced by an implicit assumption that the global minimizers of $SSTRESS()$ and $STRESS()$ are qualitatively
similar and by the aforementioned assumptions involved in deriving \( STRAIN() \) from \( SSTRESS() \).

We will briefly comment generally on these assumptions below:

a. **Replacing \( STRESS() \) by \( SSTRESS() \).** Roughly speaking, if the target distance matrix, \( L \), is sufficiently close to being consistent, then it seems likely that any differences between global minimizers of \( STRESS() \) and \( SSTRESS() \) would be insignificant. If \( L \) includes significant violations of the triangle inequality so that it is not ‘close’ to being a consistent distance matrix, then it would seem appropriate to apply methods (e.g. metrization) for refining \( L \) so that it would be ‘closer’ to being a consistent distance matrix.

b. **Spring constant as a product of two terms.** Even with the restriction that the spring constants for an atom pair are a product of a term for each of the two atoms, there is still seems to ample flexibility in defining \( \alpha \). Also, the symmetry of spring constants defined in this way is desirable.

c. **Origin is estimated center of mass.** Choosing the origin to be the estimated center of mass is convenient. But, the choice of origin has no effect on the substance of the results.

d. **\( L \) is an estimate of \( D(x) \).** This is an implicit assumption of the DMI approach. The quality of this estimate will have an effect on the resulting estimates.

e. **Estimated atomic distances from the center of mass are approximately equal to the actual atomic distances from the center of mass.** If \( L \) is a consistent distance matrix, the estimated atomic distances from the center of mass will be equal to the actual atomic distances from the center of mass. In general, the estimated-center-of-mass calculation is robust in that the effect of errors in the distance matrix have a dampened effect on the estimated center of mass.

2. There are two reasons why the quadratic approximation assumption may lead to problematic AA-DMI-ENM trajectories. First, the quadratic approximation itself could be inappropriate. Secondly, the iterative nature of the method for generating a DMI-ENM trajectory could lead to an
eventually significant accumulation of errors. The latter source seems more likely to be the cause of the problematic $C7 \rightarrow C5_\beta$ trajectory of Figure 5.10b since wild fluctuations only occur in the last few snapshots of the trajectory. The DMI-DG method is not iterative, so the order of the generation of DMI-DG snapshots is not important. The DMI-DG is reversible in the sense that there is no directionality associated with a DG trajectory. On the other hand, a DMI-ENM trajectory from conformation A to conformation B will not, in general, be identical to a DMI-ENM trajectory from conformation B to conformation A.

3. In the examples we have presented, the AA-DMI-DG trajectories were generated using nearly consistent target matrices. Based on the experience with the alanine dipeptide, AA-DMI-DG trajectories generated using a poor set of target matrices are generally more realistic than AA-DMI-ENM trajectories using the same set of target matrices, but can sometimes be problematic (data not shown). The selection of the set of target matrices are important part of any AA-DMI method.

4. One possible modification to the methods described in this work involves modification of target distance matrices that are not consistent. Metrization is a method described in [Hav1998] for modifying an inconsistent distance matrix to make it consistent or more nearly consistent. Metrization has been developed and applied for the purpose of generating single conformations, not trajectories. Because we are interested in generating trajectories, some modifications to the metrization method of [Hav1998] might be needed to apply metrization to the generation of an AA-DMI trajectory, but metrization methods for sets of target distance matrices could be a useful enhancement to an AA-DMI method.

5. It is interesting to compare the plots in Figure 5.12 for the DG trajectory and the DG-LCEM trajectory. All of the items of the DG overlay plots are highly correlated with their counterparts in the DG-LCEM overlay plots except the potential energy in second of the four plots (shown in the upper right subplots). This is the desired effect of LCEM. The essential features of the conformation are not significantly changed, but the potential energy of some snapshots is
significantly lower for the DG-LCEM trajectory. This is also a reminder that even though overall the DG trajectory using cubic spline interpolation for target matrices closely approximates an MDS trajectory, the method doesn't take potential energy into account and differences that are seemingly minor spatially can be significant in terms of potential energy.

6. The use of cubic spline interpolation in generating target distance matrices has been emphasized when knowledge of some intermediate conformations is available. It was also mentioned that linear spline interpolation could also be employed. The more intermediate conformations are available, the more likely it is that linear spline interpolation would be effective. Even if the linearity assumption is not satisfied over the entire time interval of the transition, on a subinterval, a line becomes a better approximation to a curve as the interval of the approximation becomes shorter.

7. AA-DMI can be applied to generate trajectories that correspond to conformational transitions. One important potential application is for conformational transitions for which AA-MDS trajectories are not feasible due to limitations on system size and the length of the time interval of the simulation. Another important potential application is generation of generate initial MS BV-AA-MDS trajectories. Close inspection and careful analyses and comparison of AA-DMI trajectories with existing BV-AA-MDS solutions trajectories could be helpful in determining useful AA-DMI approaches for situations in which conformational transitions are not available and, therefore, in the author's opinion represent an important opportunity to learn about the possibilities and limitations of AA-DMI.

8. In the author's opinion, an important potential application for AA-DMI is to create trajectories when experimental data for several intermediate conformations of a conformational transition are available. This type is sometimes available from time-resolved x-ray crystallography experiments or other similar types of experiments. When this type of data is available, there is better chance for generating meaningful and nearly consistent target matrices. The AA-DMI trajectories
generated from these types of experiments could be used to generate initial MS BV-AA-MDS trajectories.

5.5 Summary

All-atom distance matrix interpolation (AA-DMI) methods are methods for generation of trajectories that correspond to conformational transitions. AA-DMI methods are less computationally demanding than BV-AA-MDS, they provide atomic level detail, and they produce hidden-time trajectories. One important potential application is for conformational transitions for which AA-MDS trajectories are not feasible due to limitations on system size and the length of the time interval of the simulation. Another important potential application is generation of generate initial MS BV-AA-MDS trajectories. The example of conformational transitions in the alanine dipeptide was considered. The linearity assumption and the quadratic approximation assumption — both implicit in linear DMI-ENM — were identified. The difficulties associated with the quadratic approximation assumption can be avoided by incorporating a particular objective function, $\text{STRAIN}$, into an AA-DMI method. The difficulties associated with the linearity assumption can be alleviated by incorporating cubic spline interpolation into the generation of target distance matrices rather than the linear polynomial interpolation that results from the linearity assumption. Cubic spline interpolation is possible when approximate intermediate points of a trajectory are known or can be inferred. A method for refinement of AA-DMI trajectories by local constrained energy minimization (LCEM) for an alanine dipeptide was described and applied. Analysis is provided and directions for related future research were suggested.

5.6 References


[Cas2002] Case, D. Re: Dielectric constant for interior of proteins. AMBER Mail Reflector. 


Figures and Tables

Figure 5.1  Ball-and-stick visualization: three dihedral angles of \( \text{N-acetyl-N'}\text{-methylalaninamide} \)

Ball-and-stick visualization: \( \phi, \psi, \) and \( \theta \) dihedral angles of \( \text{N-acetyl-N'}\text{-methylalaninamide} \). Figure 1 of [Bol2000]. The C-N-C\( ^\alpha \)-C backbone dihedral angle (\( \phi \)) and the N-C\( ^\alpha \)-C-N backbone dihedral angle (\( \psi \)) are labeled. Another dihedral angle of significance in transitions between C7\text{ax} and C7\text{eq} potential wells is O-C-N-C\( ^\alpha \) and it is labeled as \( \theta \) here. \( \theta \) should not be confused with the another type of backbone dihedral angle C\( ^\alpha \)-C-N-C\( ^\alpha \) which is commonly labeled as \( \omega \).
Figure 5.2 $\varphi/\psi/\theta/DME/U$ plots: linear AA-DMI-ENM with uniform weights

Linear AA-DMI-ENM with uniform weights
On the vertical axis, $\varphi$ (black), $\psi$ (magenta), $\theta$ (dotted gray) all shown in degrees, $U(x(t_k))$ (dashed dotted red) shown in kcal mol$^{-1}$, and $100 \times DME(x^{ENM}(t_k),x^{ENM}(t_k))$ (dashed green). On the horizontal axis, the conformation index number, $k$, is shown.
Figure 5.3  $\phi/\psi/\theta$ /DME/U plots : low energy solution trajectories

From left to right, top to bottom: Figures a (C7$_{eq}$-C6), b (C7$_{eq}$-C5$\beta$), and c (C7$_{eq}$-C7$_{ax}$)

Low energy solution trajectories
On the vertical axis, $\phi$ (black), $\psi$ (magenta), $\theta$ (dotted gray) all shown in degrees, $U(x(t_k))$ (dashed dotted red) shown in kcal mol$^{-1}$, and $100\times$DME($x^{ENM}(t_k), x^{ENM}(t_k)$) (dashed green). On the horizontal axis, the conformation index number, $k$, is shown.
Dynamic distance distribution plots: $C_{7\text{eq}} \rightarrow C_6$ transition
Figure 5.5 DDD plots: approximate MDS solution: \( C_{7_{eq}} \rightarrow C_{5_{\beta}} \) transition

Dynamic distance distribution plots: \( C_{7_{eq}} \rightarrow C_{5_{\beta}} \) transition
Figure 5.6  DDD plots: approximate MDS solution: $C_7_{eq} \rightarrow C_7_{ax}$ transition

Dynamic distance distribution plot

Dynamic distance distribution plots: $C_7_{eq} \rightarrow C_7_{ax}$ transition
Figure 5.7  DDD plots: approximate MDS solution II : $C_{7\text{eq}} \rightarrow C_{7\text{ax}}$ transition

Dynamic distance distribution plots : $C_{7\text{eq}} \rightarrow C_{7\text{ax}}$ transition
Figure 5.8  DDD plots: MDS solution \( C_{7eq} \rightarrow C_{7ax} \) transition

Dynamic distance distribution plots: \( C_{7eq} \rightarrow C_{7ax} \) transition
Figure 5.9  DDD plots: linear DMI-ENM : $C_{7_{eq}} \Rightarrow C_{7_{ax}}$ transition

Dynamic distance distribution plots : $C_{7_{eq}} \Rightarrow C_{7_{ax}}$ transition
Figure 5.10  $\phi/\psi/\theta$/DME/U plots : nonlinear AA-DMI-ENM with target distances from a BV-AA-MDS solution trajectory

Subfigures a-l, from left to right, then top to bottom; Row 1: a-d; row 2: e-h, row 3: i-l

Nonlinear AA-DMI-ENM with target distances from actual trajectories
On the vertical axis, $\phi$ (black), $\psi$ (magenta), $\theta$ (dotted gray) all shown in degrees, $U(x(t_k))$ (dashed dotted red) shown in kcal mol$^{-1}$, and $100 \times$ DME($x_{ENM}(t_k)$) (dashed green). On the horizontal axis, the conformation index number, $k$, is shown. Note the wildly problematic subplots b, i, and l.
Figure 5.11  $\phi/\psi/\theta$/DME/U plots: nonlinear AA-DMI-DG with target distances from a BV-AA-MDS solution trajectory

Subfigures a-l, from left to right, then top to bottom; Row 1: a-d; row 2: e-h, row 3: i-l

Nonlinear AA-DMI-DG method with target distances from distances from BV-AA-MDS solution trajectories  On the vertical axis, $\phi$ (black), $\psi$ (magenta), $\theta$ (dotted gray) all shown in degrees, $U(x(t_k))$ (dashed dotted red) shown in $kcal\ mol^{-1}$, and $100\times DME(x^{ENM}(t_k),x^{ENM}(t_k))$ (dashed green). On the horizontal axis, the conformation index number, $k$, is shown.
Figure 5.12  $\phi/\psi/\theta/DME/U$ plots : $C7_{eq} \rightarrow C7_{ax}$ transition
DG and DG-LCEM trajectories

$C7_{eq} \rightarrow C7_{ax}$ transition:
On the vertical axis, $\phi$ (black), $\psi$ (magenta), $\theta$ (dotted gray) all shown in degrees, $U(x(t_k))$ (dashed dotted red) shown in kcal mol$^{-1}$, and $100 \cdot DME(x_{ENM}(t_k) , x_{ENM}(t_k))$ (dashed green). On the horizontal axis, the conformation index number, $k$, is shown.
Figure 5.13  DDD plots: overlay \( C_7^{eq} \rightarrow C_7^{ax} \) transition

Dynamic distance distribution

- 5 point 'min to min' BVP
- set #09
- red BV-MDS solution
- green DG spline solution
- black DG-LCEM spline solution
- magenta spline curve
Figure 5.14  DDD plots: zoom overlay : $C_{7\text{eq}} \rightarrow C_{7\text{ax}}$ transition

Dynamic distance distribution

- 5 point 'min to min' BVP
- set #09
- red BV-MDS solution
- green DG spline solution
- black DG-LCEM spline solution
- magenta spline curve

$C_{7\text{eq}} \rightarrow C_{7\text{ax}}$ transition
6 SUMMARY AND FUTURE DIRECTIONS

In this final chapter, a brief summary and an assessment of future directions are provided.

6.1 Summary

Conformational changes in proteins and other biomolecules have important consequences for many biological processes and biochemical pathways. Further understanding of the dynamics of conformational changes may lead to further progress in biological research. Simulations may lead to new hypotheses about pathways and mechanisms and may also be a platform for further testing of theories and ideas. It is important that the results from simulations be compared with results from experiments and other simulations to encourage refined simulation models and to pursue their validation. Advances in experimental methods will provide further opportunities for validation.

An all-atom molecular dynamics simulation (AA-MDS) generates a trajectory that satisfies Newtonian equations of motion with accuracy that depends on the mesh, the numerical method used for solving IVP’s, and rounding errors resulting from finite precision arithmetic and which can vary depending on the computer platform being used. The utility of an AA-MDS will vary to the extent that the force field and other simulation parameters provide a useful representation of the physical system being modeled. While AA-MDS has limitations and shortcomings, it is generally accepted that for appropriate choices for mesh, numerical method, and computer platform and for appropriate modeling considerations, AA-MDS trajectories will provide detailed and representative dynamics of a physical system. Because of this consensus view, AA-MDS holds an important place in the realm of molecular modeling and scientific research. For some particular conformational transitions of particular biomolecules, an ensemble of AA-MDS trajectories can be attained. But, in general, the limitations and shortcomings have a significant impact on this application of AA-MDS.

An AA-MDS trajectory can be viewed as a mathematical model in which the trajectory is a numerical solution to an initial value problem (IVP) or is a numerical solution to a boundary value problem.
(BVP). Historically, for most applications of AA-MDS, the former view has been preferred either for modeling reasons or for reasons of convenience or both. With the study of conformational transitions in mind, the boundary-value approach to AA-MDS, which has been described in this thesis, can take advantage of information about the desired ending structure for a simulation. Among different numerical methods for solving boundary value problems for AA-MDS, the multiple shooting methods for BV-AA-MDS that are described in this dissertation could be characterized as complementary to methods like finite difference methods and least-action methods such as the stochastic difference equation method. The transition path sampling approach to BV-AA-MDS could be described as a shooting approach, but this approach seems most useful to create an ensemble of trajectories when a solution trajectory has already been obtained and also when a reaction pathway has already been determined. Also, the number of subintervals for this approach appears to be limited to one or two and the method for updating velocity parameters on each iteration is different than those using in the shooting methods described in this dissertation. Another important aspect of the multiple shooting approach is that the approach can likely be efficiently applied in a parallel or distributed environment. The IVP trajectories from the different subintervals could be obtained from separate processors.

In Chapter 2, we observed that the conformation of an AA-MDS trajectory assumed at a point in time exhibit a sensitivity to small changes in initial conditions that increases exponentially as the different between the point in time and the initial time increases. This property has been referred to in literature as Lyapunov instability, but it seems to be more a result of the lack of asymptotic stability. Focusing on the stability of equilibrium points, we have also argued that equilibrium points in phase space corresponding to local minima of a potential energy surface are Lyapunov stable in theory and due to energy conserving properties of numerical methods like the velocity Verlet algorithm can exhibit this type of stability for long time periods in practical applications as well. This implies that for initial conditions that are measured perturbations from the equilibrium point of phase space, the system will remain within some measurable distance of that equilibrium point. The apparent inconsistency between the analysis here and other literature can probably be resolved by noting that in most applications of initial value (IV-) AA-MDS, the
initial perturbation is high enough so that the measurable distance from the equilibrium point is quite large. And, the asymptotic effect of the exponentially increasing sensitivity to initial conditions is that when the system is subject to a perturbation of initial conditions, the particular point in accessible phase space at a given point in time become random as the given point in time goes to infinity. Accessible phase space is theoretically determined by the initial perturbation and particular equilibrium starting point. If accessible phase space encompasses a wide sample of conformations, the effect of sensitivity to small changes in initial conditions will be significant.

For simulation of conformational transitions over longer time periods, the lack of asymptotic stability suggests that single shooting methods cannot be expected to be effective. For longer time periods, finite difference methods with a coarse mesh can only be expected to produce approximate trajectories and finite difference methods with a mesh sufficiently dense to produce accurate trajectories can involve a prohibitively large number of parameters. Like single shooting time intervals, multiple shooting subintervals must be limited in length in order for the method to be effective since a multiple shooting method involve revision to initial velocities at shooting points based on results from previous iterations. But, like the choice of mesh for finite difference methods, the choice of the mesh of multiple shooting points must be limited in density to avoid an excessive number of parameters. It has been argued here that multiple shooting methods involve a reasonable middle ground between single shooting and finite difference methods, thereby limiting the problems associated with applying those methods over long time intervals. The scope of applicability of multiple shooting would be broadened if parameter reduction methods, like those employed for single shooting in Chapter 4, could be effectively employed for multiple shooting. In Chapter 3, a multiple shooting approach for finding rare, low-energy transition trajectories between local minima was described. In Chapter 4, an approach for approximating potential energy wells by bounds on interatomic distances was introduced as well the aforementioned approach for parameter reduction. In Chapter 5, promising approaches to all-atom distance matrix interpolation methods that resolve problems associated with the quadratic approximation of current methods and address problems associated with the commonly applied linearity assumption were provided. All-atom distance matrix
methods have been effectively used to generate initial trajectories for multiple-shooting methods that are applied find AA-MDS trajectories corresponding to conformational transitions. There is a potential application of the distance matrix interpolation methods introduced here for modeling conformational transitions over long time intervals in large systems that are unreachable by AA-MDS. Considering the work presented in these three chapters, it would seem worthwhile to consider applying the multiple shooting methods and distance matrix interpolation methods described here to larger molecules.

6.2 Future directions in applications

Below are some more detailed comments on possible future directions in applications.

- A primary goal would be to attempt to apply our multiple shooting methods to a significantly larger system. Initial value approaches to all-atom molecular dynamics simulation have been used to simulate folding of a 16-residue, 253-atom, β-hairpin of immunoglobulin-binding Protein G. (The PDB identification code, or PDB id, is 1GB1.) This β-hairpin or a similarly sized molecule would seem to be a significant but reasonable increase in size. We have done some preliminary testing for this molecule using a MATLAB implementation of the AMBER all-atom force field. Consider applying the multiple shooting method with a full set of $6nN = 6 \times 253 \times N = 1518 \times N$ parameters to a boundary value problem with a full set of $6n = 6 \times 253 = 1518$ boundary conditions. For an iteration of the multiple shooting method, to compute the Newton step, one needs to compute an almost-block-diagonal $1518N \times 1518N$ Jacobian and solve a resulting linear system of $1518N$ scalar equations in $1518N$ unknowns. There is not much computational burden for computing the Jacobian. And, using the approach described in Chapter 7 of [Sto2002], one can solve the linear system to update the $1518N$ parameters by doing one $6n \times 6n$, or $1518 \times 1518$, linear system solve which is of order $O(n^3)$. There is not much computational burden in doing a dense linear system solve with these dimension. This linear system solve is followed by a series of calculations involving matrix multiplication which is of order $O(n^3)$. 
• It is anticipated that a primary focus of future work is all-atom molecular dynamics simulation, but there is the possibility of applying multiple shooting methods to coarse-grained, or reduced, models of proteins. In [Mal2005], interesting results and physically meaningful results are obtained from molecular dynamics simulations of prion-like proteins performed using a coarse-grained lattice model with a potential energy function that was empirically derived from protein structure databases.

• It would be of interest to apply the distance matrix interpolation methods developed in this work to study conformational transitions for larger proteins over longer time periods. Perhaps, it would be useful to include application of these methods to cases previously studied using coarse-grained DMI-ENM (e.g. [Kim2002a], [Kim2002b], [Kim2003]). Examples include toy models of planar motions (elongation, shear, hinge bending, breathing, ligand binding, Holliday junction formation); lactoferrin (PDB id’s: 1LFG, 1LFH); lac repressor headpiece: (PDB id’s: 1LCC; 1LCD); lactate dehydrogenase: (PDB id’s: 1LDM; 6LDH); citrate synthase: (PDB id’s: 4CTS; 1CTS); and 16S Ribosomal RNA. Other molecules for which use of distance matrix interpolation has not been reported but which exhibit interesting and important conformational changes, include adenylate kinase, photoactive yellow protein, RAS, calmodulin, and prion.

6.3 Future directions in methods and analysis

Some possible future direction in terms of methods and analysis are outlined below:

• In Chapter 4, a method for parameter reduction was presented for the multiple shooting method with one subinterval, i.e. single shooting. A primary objective for future work is to develop a method for parameter reduction for multiple shooting with more than one subinterval.

• The timing of conformational transitions can be an important aspect of the role that the transitions play in biological processes. In Chapter 4, it was shown that the realization of a particular conformational transition of an alanine dipeptide was disproportionately impacted by the magnitude
of initial velocities in a small subset of directions corresponding to eigenvectors for vibrational modes of low frequencies. A closer examination of these eigenvectors could be helpful in understanding the mechanism for a particular transition. Another possibility is to explore and identify possible relationships between external forces and the likelihood of conformational transitions in proteins. It may be that a particular external force on a particular atom or subset of atoms could be correlated with one or more of the important eigenvectors for a conformational transitions or more directly correlated with an actual conformational transition. Multiple shooting and distance matrix interpolation methods may be useful in the study of relationships between external forces and conformational transitions in proteins.

- It could be beneficial to develop a greater understanding of the relationship between the stochastic difference equation approach and the multiple shooting approach by analytical means and by the study of numerical experiments for small model systems.

- It could be useful to consider different ways to model solvent effects, to understand more about Brownian dynamics simulation, and to consider the relationship between these areas of research and the work presented here.

- With respect to numerical methods, there are many aspects of this work that could benefit from a closer look. As an example, both of the global convergence methods that we use involve calculation of the Jacobian. For the 253-atom β-hairpin, this calculation is not problematic since elements necessary to compute the Jacobian can be computed step-by-step as the IVP’s are solved. However, for much larger systems, this approach could become problematic. So, other less expensive global convergence methods could be considered.
ACKNOWLEDGEMENTS

It has been a privilege and an enjoyable experience to be a part of the research groups of Dr. Zhijun Wu and Dr. Robert Jernigan here at Iowa State. I am thankful to be a part of these groups and consider myself fortunate to have Dr. Wu and Dr. Jernigan as advisors. I thank Dr. Ron Elber for providing me with an opportunity to visit his research lab. I thank my other committee members, Dr. Kai-Ming Ho, Dr. Howard Levine, and Dr. Michael Smiley for helpful comments and willingness to participate in my committee. I thank all my committee members for their helpful advising, for being interested in and supportive of my research efforts, and for sharing with me some of their accumulated wisdom on the arts of researching, writing, and presenting. I would also like to thank the following: the NSF-IGERT program, the Department of Mathematics, the Program in Bioinformatics and Computational Biology, Pioneer Hi-Bred International and in particular the Bioinformatics Group there including Lane Arthur and former member Jean Peccoud. I am thankful to family and friends beyond Iowa State for their interest and support. I will have many good memories of Iowa State including the activities and people of the aforementioned research groups of Dr. Wu and Dr. Jernigan; classes that I attended; classes that I taught; Department of Mathematics; Program in Bioinformatics and Computational Biology; ISU/UI Bioinformatics Workshops; ISU Bioinformatics and Computational Biology Summer Institute; Dr. Volker Brendel’s running and research group; Dr. Xun Gu’s research group; Dr. Dan Nettleton’s research group; Beyer Hall pickup basketball; Water Polo Club; Ames/ISU Ice Arena pickup hockey; Iowa Martial Arts Group; Descarga; Dub H; Music Hall; Memorial Union; Parks Library; and others. I also enjoyed having the opportunity to represent Iowa State at intercollegiate water polo club tournaments and at regional and international conferences on bioinformatics, computational biology and applied mathematics.