**Crystal Molecular Dynamics**

**by**

**Paweł A. Janowski**

**A dissertation submitted to the**

**Graduate School of New Brunswick**

**Rutgers, the State University of New Jersey**

**in partial fulfillment of the requirements**

**for the degree of**

**Doctor of Philosophy**

**Graduate Program in Chemistry and Chemical Biology**

**and**

**Graduate Program in Computational Biology and Molecular Biophysics**

**Written under the direction of**

**Prof. David A. Case**

**and**

**Prof. Darrin M. York**

**and approved by**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**New Brunswick, New Jersey**

**August, 2015**

# Abstract of the dissertation

**Improved Molecular Dynamics and Macromolecular Crystallography through Simulations of Biomolecular Crystals**

**By Paweł A. Janowski**

**Dissertation Directors**

**David A. Case Ph.D and Darrin M. York Ph.D**

We present a broad effort at the development of crystal simulation methodology and its application to benefit both macromolecular crystallography and molecular dynamics methods. Crystallography is the current method of choice for structural determination of biomolecules, but it is hampered by the inherently time and space averaged nature of the experiment as well as methodological limitations that do not sufficiently account for the heterogeneous and dynamic nature of crystals. Molecular dynamics has proven itself as a method capable of probing the physics and chemistry of biomolecules on an atomic scale, but requires continued development of the underlying force field parameters to more accurately reproduce observables. Our effort has focused on developing the framework for molecular dynamics simulations of biomolecular crystals. We first present our methodology for performing crystal simulations and show how it is applied first to simple peptide crystals and then to increasingly complex biomolecular systems. Next we demonstrate the utility of crystal simulations for validation of molecular dynamics methods through two case studies of the biophysics of enzyme reactions. Finally we demonstrate the improvement to crystallographic methods that can be gained by incorporating molecular dynamics methods. Our work is of great benefit to both the molecular dynamics and macromolecular crystallography communities and proposes specific approaches to integrate the two fields for the benefit of both.

# Acknowledgements

I thank the Lord for the beautiful gift of His Creation, for the flowers, the mountains, the creatures and the fascinating molecular mechanisms of life that testify to His love. Thank you for the gift of reason, free will and desire of the Good, the True and the Beautiful that drives us in all our pursuits. May we one day attain the fullness of Love.

I thank Maria my wife who accompanied me throughout this path in person and in spirit. Thank you for your care, your patience, your inspiration. We have truly walked the past five years together in friendship. May we walk many more!

I thank my parents, Jolanta and Andrew, for their love in raising me to be the person I am. Thank you for teaching me what is important in life and what is not. I am forever indebted in the bond of filial love.

I thank my family and my friends for all your kindess, your love, all the good times spent together that will forever form part of the treasure of my memories. Without those wonderful times shared together, I would have been hard pressed to keep my sanity along the way.

I thank all my colleagues, my lab mates, my teachers, all the wonderful scientists I have had the pleasure to work with. Especially Prof. Karsten Krogh-Jaspersen and Prof. Krzysztof Lewiński for transmitting the beauty of science through their lectures. Thank you for showing me how exciting science is!

I thank all of the good people I have met. Each encounter with each one of you enriches both of us. Only in relationship with others does man become truly man. Never stop being who you are. I love you all.

I thank my collaborators with whom I have had the enormous fortune of interacting over the course of the last five years. I have learned so much from all of you! Thank you for all of your support and encouragement. In no particular order: David Cerutti, Nigel Moriarty, James Holton, Chunmei Liu, Paul Adams, Nathaniel Echols, Pavel Afonine, Thomas Terwilleger, Jane Richardson and the entire Phenix team, Greg Warren, Brian Kelley, Anthony Nicholls, Timothy Giese, Jason Swails.

Finally, and in a special way, I thank my advisors, Prof. David Case and Prof. Darrin York. You have challenged and inspired me constantly and at the same time been caring guides along the path. I have learned so much from you, not just about science but about what it means to be good human beings. I will remain forever grateful.

To the loving memory of my Father who would have wanted to be here but is even closer than we can imagine.

# Contents

[Abstract of the dissertation ii](#_Toc423711142)

[Acknowledgements iv](#_Toc423711143)

[Contents vi](#_Toc423711144)

[Abbreviations used vii](#_Toc423711145)

[Section I. Introduction 1](#_Toc423711146)

[Chapter 1. Introduction and background 1](#_Toc423711147)

[Section II. Developing molecular dynamics of crystals. 16](#_Toc423711148)

[Chapter 2. Fav8 1 16](#_Toc423711149)

[Chapter 3. Fav8 2 17](#_Toc423711150)

[Chapter 4. 4lzt 18](#_Toc423711151)

[Chapter 5. DNA/RNA 19](#_Toc423711152)

[Section III. Applications of molecular dynamics of crystals 20](#_Toc423711153)

[Chapter 6. Hairpin 20](#_Toc423711154)

[Chapter 7. RnaseA 21](#_Toc423711155)

[Section IV. Improved crystallographic methods through crystal molecular dynamics 22](#_Toc423711156)

[Chapter 8. AFITT 22](#_Toc423711157)

[Chapter 9. Phenix-Amber 23](#_Toc423711158)

[Bibliography 24](#_Toc423711159)

# Abbreviations used

MD – Molecular dynamics;

BX – Biomolecular crystallography;

FT – Fourier transform;

Å – Ångstrom;

PDB – Protein Data Bank;

NVE – microcanonical ensemble;

NVT – canonical ensemble;

NPT – isobaric-isothermal ensemble;

ps – picosecond;

fs – femtosecond;

ns – nanosecond;

μs – microsecond;

ms – millisecond;

PME – particle mesh Ewald;

PBC – periodic boundary conditions;

# Introduction

## Introduction and background

When, during my initial visit to Rutgers, Prof. David Case first mentioned the idea of improving crystallography through molecular dynamics of crystals, I felt a tinge of excitement. I had studied crystallography for two semesters during my undergraduate coursework at Jagiellonian University in Krakow. Lectures were eloquently delivered by one of the best teachers I’ve ever had, Prof. Krzysztof Lewiński. However, despite all my effort I could not grasp the essence of how a seemingly random pattern of dots on a sheet of paper could be turned into a three dimensional model of a biomolecule. I liked crystallography, but I also respected it and I feared it because I felt like there was something powerfully beautiful and mysterious about it. So when Dr. Case floated this idea of molecular dynamics of crystals I was excited: here was a chance to make up for my previous failing, to finally come to understand crystallography or to die trying. And to do that by using the molecular dynamics that I wanted to focus my Ph.D. studies on… it was the perfect project.

Thus I have happily spent the last five years focused on our effort to simulate biomolecular crystals with molecular dynamics. The original question we asked ourselves was simple: what can we learn from molecular dynamics of crystals? And this was quickly reformulated into the following four overarching questions that form the focus of this work:

1. What is the best way to carry out molecular dynamics of biomolecular crystals?
2. How can we use crystal simulations to improve molecular dynamics methods?
3. How can we use crystal simulations to improve crystallography methods?
4. What can we learn about real crystals from our simulations of crystals?

What follows is a brief introduction to the methods of crystallography and molecular dynamics, with special emphasis on aspects that relate directly to our work. We then discuss the goals and specific aims of this research and present the general organization of the dissertation before moving on to a presentation of the work in subsequent chapters.

### Crystallography background

Crystallography is a biophysical technique used to probe the three-dimensional distribution of atoms in molecules by analyzing the diffraction pattern of electromagnetic radiation on a crystal. As the name implies, crystallography requires that billions of copies of the molecule of study arrange themselves in a regular repeating array which is, by definition, a crystal. When used to study the structure of biomolecules, the method is referred to as macromolecular crystallography (MX). The fact that protein molecules can form crystals has been known for almost 150 years. In general, crystal formation of biomolecules is promoted by slowly removing solvent from a solution of the protein of study. If the solvent is removed too quickly or if the solution is not of the required purity, the protein molecules will precipitate out of the solution and form an amorphous powder. However if the solution becomes supersaturated slowly the molecules may pack themselves in a regularly repeating array held together by non-covalent chemical interactions in a way that minimizes the overall energy of the solute. Finding the exact conditions under which a given biomolecule crystallizes can be very challenging and in many cases constitutes the crux of the crystallographic method.

Once crystallized, the regularly repeating array of the crystal acts as a diffraction grating when light is shined upon it. Diffraction refers in general to the physical behavior of waves as they impact objects or slits. Etimologically, the term was coined by Francesco Maria Grimialdi in 1660 and comes from the Latin diffringere meaning “to break up into pieces”. In particular a regularly spaced array of slits or objects will cause the waves scattered off each object to interfere with each other. Wave crests lining up leads to constructive interference resulting in waves of higher amplitude, whereas when crests and troughs mix, destructive interference results in low amplitudes. Because of the dual nature of electromagnetic radiation, when light shines on diffraction grating it behaves like a wave and interference leads to the formation of bands (in the case of a one-dimensional diffraction grating) or spots (in the case of a two-dimensional diffraction grating). James Gregory’s observation of the diffraction pattern of light shining through a bird feather in the late 17th century constituted the discovery of the first diffraction grating.

A crystal is a repeating array of objects and thus can naturally act as a diffraction grating. However, because the wavelength of visible light is much larger than the typical spacing between array planes in molecular crystals, the diffraction of light on molecular crystals is not observed. The breakthrough moment for crystallography came in 1912 during a conversation between Paul Peter Ewald and Max van Laue, when van Laue suggested that x-rays (discovered in 1895 by Wilhelm Roentgen) might have a shorter wavelength that would allow their diffraction on crystals to be observed. In 1912 van Laue recorded the first ever x-ray diffraction pattern on a copper sulfate crystal. Shortly thereafter the father-son pair of William Lawrence Bragg and William Henry Bragg formulated the law that describes the diffraction of x-rays on a crystal. The first diffraction pattern from a protein crystal was obtained by John Desmond Bernal and Dorothy Hodgkin using pepsin, and the first three-dimensional structure of a protein molecule solved using x-ray crystallography was myoglobin in 1958 by John Kendrew. Van Laue, the Braggs and Kendrew all received Nobel Prizes for their work. In all thirteen Nobel Prizes have been award for work on or using crystallography.

The raw experimental data obtained in a crystallography experiment is a diffraction pattern (fig 1). This pattern is obtained as a beam of x-rays is focused on a crystal and the x-ray photos scatter (diffract) off the electron clouds of the atoms that make up the crystal. For the work presented here it is crucial to understand that the diffraction pattern is not obtained in a single instant from single x-rays scattering off the crystal. Rather it is obtained over a significant period of time usually ranging from a few up to about 30 minutes. The diffraction spots themselves require the constructive interference of a enormous number of x-rays to be observed. Furthermore the xrays themselves diffract off the billions plus molecules that make up the crystal. Thus one can say that crystallography is truly a time and space averaged experimient.

The diffraction patterned obtained in the crystallography experiment contains to essential pieces of information. The first of these is the location and spacing of the diffraction spots. The spots appear on the vertices of an array called the reciprocal space lattice which is a mathematical construct directly related to the parameters of the real space lattice. The real space lattice is the lattice of the crystal being studied. The appearance of diffraction spots can be described via the equation that is know as Bragg’s Law (named for the younger of the two Braggs mentioned above). Essentially, diffraction spots can only form in locations where the x-rays arrive in phase (in more simple language, where the crests and troughs of the arriving x-ray waves are lined up with each other). The condition for this to happen is that the distances of the paths that all the arriving x-rays travelled must all differ by an integral number of wavelengths (wavelength is the distance from one crest of the wave to the next) of the x-rays. This is presented schematically in Figure xx. The resulting description of the formation of diffraction spots is Bragg’s Law:

|  |  |  |
| --- | --- | --- |
|  |  | (Eq. ) |
|  |  |  |

where d is the spacing between a given set of planes in the array, θ is the angle at which the x-rays impact the set of planes, *n* is a positive integer and λ is the wavelength of the x-rays. The lattice (spacing between planes) and wavelength are constant under normal experimental conditions. Thus they uniquely specify the angle at which the scattered x-rays interact constructively and form a diffraction spot. Each spot thus corresponds uniquely to a specific set of planes in the array. Furthermore the angle is inversely proportional to the spacing. In other words smaller diffraction angles correspond to larger plane spacing in the lattice. Diffraction spots closer to the center of the diffraction pattern carry information about larger-scale features of the crystal. This is the basis for the concept of resolution: usually the diffraction pattern is only measured up to a certain radius: beyond that the angle of diffraction is too large and the spots too weak to be reliably recorded. Most importantly, by accurately measuring the location and spacing of the diffraction spots, one can deduce the spacing of the crystal’s array and thus obtain the parameters of the crystal unit cell (the three box dimensions *a*, *b*, *c* and three box angles α,β,γ.

The other essential information in the diffraction pattern are the intensities of the diffraction spots. Where the location of the spots reveals the unit cell parameters of the crystal array, the intensities of the spots tell us about the actual distribution of scattering objects, i.e. atoms, within each unit cell. The intensity of wave is equal to the square of it’s amplitude:

|  |  |  |
| --- | --- | --- |
|  |  | (Eq. ) |
|  |  |  |

We know from the previous discussion and Bragg’s Law that an identical scattering object located at each lattice plane of a certain spacing *d* would produce an ideal constructive interference between x-rays and consequently a diffraction spot at angle θ. But what happens if there are additional scattering objects located between the planes (Fig 1). The x-rays scattering off these objects will arrive at the diffraction location with a phase different from that of the rays scattering from the primary object. The resulting amplitude of the x-ray wave arriving at the diffraction spot location is obtained by summing the waves diffracted of each object within the crystal unit cell. Because the objects do not all lie integral distances of the scattering plane away from each other, the resulting waves that are summed are not all perfectly in phase. This results in an attenuation of the amplitude of the resulting wave and in extreme cases (Fig…) can results in a complete disappearance of the spot. If we treat the scattering electron density in the unit cell as continuous and divide it into infinitesimal sections *dx* along the scattering vector, the amplitude of the resulting diffraction spot can be obtained by integrating the partial x-ray wave scattered by each section *dx* of electron density:

|  |  |  |
| --- | --- | --- |
|  |  | (Eq. ) |
|  |  |  |

(Eq. 3) is presented for the one dimensional case but the generalization to three dimensions is straightforward. Each partial wave has an amplitude proportional to the electron density at *x* but with a phase relative to *x=0* of *2πhx*. The integration is performed over the unit cell vector *h* and the position *x* are described in fractional coordinates. ρ(x) is the electron density at position *x.* ***Fh*** is called the structure factor and is a wave described by an amplitude  and a phase . The intensity of the diffraction spot is related to the structure factor amplitude via (Eq. 2) As it turns out, this equation is equal to the mathematical transformation known as the Fourier Transform (FT). If the electron density is presented discontinuously as a set of scattering points(atoms) we obtain the discreet form of the structure factor equations:

|  |  |  |
| --- | --- | --- |
|  |  | (Eq. ) |
|  |  |  |

which we have now presented in three dimensions. *f* is the scattering contribution of atom *i* in the scattering direction corresponding to reflection ***h***. Conversely if we sum over each one of the diffracted waves (at each diffraction spot), we obtained the scattering electron density:

|  |  |  |
| --- | --- | --- |
|  |  | (Eq. ) |
|  |  |  |

Here again we present the one-dimensional form for pedagogical purposes. The summation is over all the diffraction spots of order *h*. This equation corresponds to the form of the inverse Fourier Transform and is the mathematical inverse of the (Eq. 3). Thus we arrive at one of the fundamental concepts of x-ray crystallography: the electron density of the crystal unit cell is the inverse FT of the diffraction pattern.

Now let us examine what is needed to calculate the electron density. (Eq. 5) states that we need to perform a summation over each diffraction spot. For each spot we need the amplitude and phase of its corresponding structure factor. The amplitude is readily obtained as the square root of the intensity measured in the experiment, but unfortunately there is no information about the phase. This is known as a phase problem. Many ingenious (and difficult to implement) methods exist to tackle the phase problem. Here let it suffice to say that if a sufficiently good estimate of the phases is obtained from which a sufficiently good estimate of the electron density can be calculated, then one can move on to the next part of the process, refinement, that is of much greater concern to us in the present work. In practice, the great majority of biomolecular structures are solved today by a technique called molecular replacement where a sufficiently good initial estimate of the electron density and phases is obtained by comparison to another similar molecule whose structure is already known.

Supposing that a fairly good estimate of the structure of the molecule has been obtained, one can move on to the next stage in the crystallography process which is called refinement (structural refinement, crystallographic refinement). Let us summarize what information we have at this stage. From the experiment we have the amplitudes of all the structure factors. If we also had the phases we could be able to calculate the electron density by (Eq. 5), but we usually don’t have the phases. One the other hand we have an estimated structure of the molecule. This is referred to as a model and usually consists of a position relative to the crystal unit cell for each atom that we know makes up the molecule we are studying. From these atomic positions we can calculate the overall all electron density of the model (the electron density is modelled by some mathematical function of the atom type. The functions most commonly used today are the Cromer-Mann Gaussian functions. ) Then from the electron density of the model we can calculate amplitudes and phases via (Eq. 3). We now have a set of experimentally measured structure factor amplitudes which are commonly referred to as Fobs, and we have a set of structure factor amplitudes calculated from the current best estimate model of the molecule, which are commonly referred to as *Fcalc*­. We can now quantify how well the proposed model accounts for the experimental data (or alternatively, how well the experimental data describes the proposed model) by comparing *Fcalc* to *Fobs*. This is usually done by a statistic known as the R-factor:

|  |  |  |
| --- | --- | --- |
|  |  | (Eq. ) |
|  |  |  |

where and are the amplitudes of the *F­­obs* and *Fcalc* set of structure factors respectively. We can envision the following process: given a starting model we calculate the R-factor. We know uses the phases obtained from the model via (Eq. 3) together with the experimental amplitudes *Fobs* to calculate an electron density via (Eq. 5). Next we adjust the atomic positions of our model to better fit the electron density calculated from *Fobs.* From there we calculate a new set of *Fcalc*and a new R-factor. If the R-factor is better (lower) than the previous one, than the new model is better than the previous model. This iterative process of calculating the electron density using phases from the model and adjusting atomic positions of the model to fit the resulting density is called refinement.

Refinement is a complex process and one could carry out the process just described by hand for a very long time and not obtain any significant improvement. Fortunately refinement can be formulated mathematically as a non-linear optimization problem and solved via one of many known mathematical algorithms. In the most basic formulation a least squares residual between the observed and calculated structure factor amplitudes is minimized:

|  |  |  |
| --- | --- | --- |
|  |  | (Eq. ) |
|  |  |  |

However, in practice this problem is often not well-defined because of the low ratio between the observed data (the set of structure factor amplitudes) and the parameters to be estimated via the optimization (the set of x,y,z coordinates of all the atoms in the asymmetric unit of the crystal) combined with the various sources of noise and error inherent in the x-ray diffraction experiment. Therefore several approaches exist to increase the data to parameter ratio. For example one can decrease the number of parameters to be refined by ignoring some set of atomic coordinates such as the hydrogens. Alternatively, one can increase the set of “observed” data by incorporating previous knowledge about the structure of molecules into the equation. For example, we know that an sp3 carbon-carbon bond should have a length of 1.54Å. This knowledge imposes a set of restraints on the final solution set of atomic positions in the molecule. Thus the residual to be minimized becomes:

|  |  |  |
| --- | --- | --- |
|  |  | (Eq. ) |
|  |  |  |

Here the x-ray term correspond to the same residual as in (Eq. 7). The chemistry (also sometimes called stereochemistry or geometry term) corresponds to the summed residual over all the restraints where r0 is the target value of the restraint and rcalc is the value of the restraint in the proposed model. *w* is a relative scaling weight that is adjusted in the refinement procedure to adjust the relative weight between the x-ray and the restraint term. The restraints used can be obtained from a variety of previously known information about the chemical structure of molecules but most commonly include knowledge of bond lengths, angles and torsions. The most popular crystal refinement programs in use today apply a set known as the Engh & Huber restraints which were derive from survey of accurate small molecule crystal structures from the Cambridge Crystallographic Database.

In practice several additional levels of complexity are present in modern refinement programs. First, the least squares formulation of the residual to be minimized is most often replaced with a maximum likelihood formulation. This allows for a statistical treatment of observation and restraint probabilities. The chemistry restraints can be incorporated as *a priori* knowledge in a Bayesian formulation. Statistical probability estimates can then be obtained on the resulting parameters. Furthermore, by incorporating this statistical knowledge a large degree of the model bias present in the calculated electron density maps due to the use of phases obtained from the model can be removed. Second, sophisticated mathematical algorithms such as the Limited memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) algorithm are implemented to optimize the residual based on gradients of its component terms. Third the equation for calculating the structure factors from the model is often more complex than the integral shown in (Eq. 3) as it includes contributions from overall anisotropy and fluctuations and from the contribution of the unmodelled bulk solvent atoms that don’t show up distinctly in the experimental electron density. Lastly, crystallographic refinement usually proceeds in stages where the refinement of the x,y,z positions of the atoms in the asymmetric unit is just one stage. Other parameters that affect the calculated structure factors are refined in the other stages. Arguably the most important of these are the B-factors. Where the x,y,z coordinates describe the mean positions of the atom in the structure, B-factors describe how that atom’s instantaneous position fluctuates around that mean. A significant portion of that oscillation can be ascribed to thermal fluctuations. Thus B-factors are often also referred to as temperature factors. B-factors can be isotropic (describing a spherical isotropic fluctuation around the mean position and leading to a single additional parameter to be refined per atom) or anisotropic (describing a three dimensional elliptical oscillation, requiring a symmetric 3x3 tensor and thus 6 additional parameters to be refined per atom). In real space B-factors act like a convolution of a Gaussian function with the electron cloud around the mean position of an atom, effectively smearing out that atom’s electron density. The equation for the FT of the electron density thus becomes:

|  |  |  |
| --- | --- | --- |
|  |  | (Eq. ) |
|  |  |  |

Importantly the isotropic B-factor is related to the physical mean displacement of the atom around its mean position by the following equation:

|  |  |  |
| --- | --- | --- |
|  |  | (Eq. ) |
|  |  |  |

Other stages within the refinement process include refinement of bulk solvent contribution, overall anisotropic scaling parameters, atomic occupancies and alternate conformations, rigid body motion and translation-libration-screw (TLS) parameters. A full completion of each of the stages of refinement is usually referred to as a macrocycle. A complete solution of a crystallographic structure usually requires many macro-cycles of refinement interspersed with stages of manual adjustment of the structure to better fit the electron density.

The end result of refinement and of the crystallography experiment in general is a complete three dimensional structure of the atoms in the molecule as it is found in the crystal of study. As of July 4th, 2015, there were 110071 biomolecular structures in the Protein Data Bank (PDB) of which 98000 had been solved by x-ray crystallography. This represents 89% of all solved biomolecular atomic structures making x-ray crystallography by far the most important contributor of data to structural biology.

### Molecular Dynamics Background

Molecular dynamics is a computational technique that aims at analyzing the internal dynamics of a physical multi-body system such as a liquid, a gas or a molecule. The was first developed by B.J. Alder and T.E. Wainright1 and independently by A. Rahman6 in the late 1950’s and early 1960’s. It was originally invented as a method to study hard sphere collisions in statistical physics, but quickly grew in its application to other fields. The first simulation of a protein was a study of bovine pancreatic tripsin inhibitor by McCammon et al.5 in 1977. Today molecular dynamics simulations are routinely performed for a wide variety of applications ranging from biophycis and chemistry to atmospheric sciences and astrophysics. Molecular dynamics can be used to obtain both a time resolve detailed view of the dynamics of the system as well as to calculate thermodynamic statistical averages over the system of study.

Molecular dynamics essentially models the system of study as a set of balls connected by springs. Dynamics of the system is obtained by applying Newtonian physics. In fact by Newton’s second law of motion we have

|  |  |  |
| --- | --- | --- |
|  |  | (Eq. ) |
|  |  |  |

where ***F*** is the force on a body and is the second derivative of the position which is the acceleration on that body produced by said force. Because acceleration is the first derivative of velocity and the second derivative of position and can be related to the former two by:

|  |  |  |
| --- | --- | --- |
|  |  | (Eq. ) |
|  |  |  |
|  |  | (Eq. ) |
|  |  |  |

given initial positions and velocities, one can integrate the acceleration at a given time to obtain new velocities and positions. By (Eq. 11) to obtain the acceleration, one needs the force, but the force is known to be minus the gradient of the potential energy of the system:

|  |  |  |
| --- | --- | --- |
|  |  | (Eq. ) |
|  |  |  |

Thus, by calculating the potential energy of a system with respect to the coordinates of the bodies that make up the system, one can take the gradient of the potential energy with respect to a specific body’s position to obtain the force acting on that particle. From there an updated set of velocities and coordinates of the body can be obtained by integrating the laws of motion (Eq. 12) and (Eq. 13). By applying this to all bodies in the system at a given time and by iterating the process over subsequent moments in time a “movie”-like trajectory of the dynamics of the system can be obtained.

We now examine how to calculate the potential energy of the system. The potential energy equation can take many forms depending on the system being studied. In the case of biomolecular systems, the most common molecular dynamics software packages in use today (Amber2, CHARMM4, NAMD, Gromacs3) use a similar potential function. In the case of Amber, which is the program used in this work, the potential energy function has the form

|  |  |  |
| --- | --- | --- |
|  |  | (Eq. ) |
|  |  |  |

The terms in the potential energy equation correspond to bond, angle, torsion or dihedral angle, Lennard-Jones or van der Waals interaction and electrostatic interaction energies respectively. The bond term makes it immediately clear why in molecular dynamics the many-body system is treated via a “beads on springs” model: the bond energy is calculated as the square of the deviation of the current bond length *r* from the ideal or target bond length *r0*  times a constant *k* which is equivalent to Hook’s law for the potential energy of a spring displaced from equilibrium. Angles and torsion angles are treated similarly with the torsion term incorporating the trigonometric function to account for a periodicity of at most 2π. The fourth term of the equation accounts for quantum repulsive and dispersive forces, sometimes known as van der Waals forces. These interactions result from the repulsion of electrons from each other as two atoms draw near to each other (why two atoms cannot overlap) and from the relatively weak attraction between atoms due to instantaneous anisotropy in the electrostatically charged electron clouds as two atoms are separated from each other. The mathematical form of this term is known as the Lennard-Jones potential and has been found to model the repulsive/dispersive interactions sufficiently well. The final term is accounts for electrostatic interactions between the charges of individual atoms in the system.

Examination of the potential function reveals what is needed to run a molecular dynamics simulation. First, one requires starting coordinates of the atoms in the system. These are necessary to calculate the distances between atoms pairs as well as bonds, angles and torsions. Sometimes the velocities are also provided but if not they can be assigned from a Boltzmann distribution at a given temperature. Second, one needs to know which atoms are connected by bonds. This allows for the summations over all atoms connected by bonds, angles or torsions as well as all remaining pairs of non-bonded atoms. This information is referred to as the topology of the system. Finally, one requires the parameters that go into the potential energy function. These include the ideal bond lengths, angle and torsion measures, the Lennard-Jones parameters for different types of atoms as well as the electrostatic charges of atoms required to calculate the Coulomb interaction. This collective set of parameters used to calculate the potential energy function given a set of atomic positions and topology is known as a force field.

The greater bulk of effort at developing and improving the accuracy of molecular dynamics simulations goes into deriving better sets of force field parameters. Force field parameters are derived by fitting simulated properties to calculations obtained through *ab initio* quantum methods or to experimental measurements of thermodynamic or spectroscopic properties. Several force fields exist in the Amber program. The most recent force field as of this writing is the ff14SB Amber force field. It is an elaboration on the ff99SB, ff10SB and ff12SB force fields that preceded it. Other available force fields include Amber ff14ipq, CHARMM36, OPLS as well as the AMOEBA polarizable force field that allows for changes to atomic partial charges as the simulation proceeds.

This generally simple outline of molecular dynamics is made slightly more complex by a multitude of enhancements mostly aimed at improving the accuracy and/or the computational efficiency of the simulation. First, a straightforward run of molecular dynamics replicates the thermodynamic microcanonical ensemble where the number of particles, volume of the system and total energy of the system are constant (NVE). However by adding computational algorithms to maintain a specified temperature or a specified pressure in the system, the canonical constant particles, volume and temperature (NVT) and the isobaric-isothermal constant particles, pressure and temperature (NPT) ensembles can be enforced. These temperature and pressure monitoring algorithms are called thermostats and barostats and the most common algorithms in use today include the Berendsen, Langevin and Monte Carlo. The NPT ensemble is of particular importance in our work as it allows the system volume to fluctuate under constant temperature and pressure as is the case with a physical crystal in experimental conditions.

Second, there are enhancements aimed at increasing the efficiency of the molecular dynamics calculation. Each cycle of calculating the potential energy and its gradients with respect to atomic positions, integrating the equations of motion and updating the atomic velocities and positions is called a step. The length of the step can be regulated by specifying the value of *t* in (Eq. 12) and (Eq. 13). The rules of statistical thermodynamics govern the rate of occurrence of events of interest at the atomic/molecular scale. Some events, such as the rotation of an amino acid side chain around a torsion angle, may occur often, on a nanosecond time scale. Other events, such as the complete folding of a protein require orders of magnitude more time, usually on the microsecond timescale. Therefore, it is desirable in molecular dynamics to simulate a length of time sufficient to be make it probable that the event of interest will occur within the simulated time window. However, one cannot simply make the time step larger: if the time step is larger than the time scale of the fastest events simulated by the force field, the integration of positions and velocities will proceed in leaps without responding in time to the effects of these events and resulting in severe instabilities in the system. Normally the fastest events observable in the simulations are bond length vibrations that occur on a femtosecond scale. Thus the time step employed in a typical simulation is usually 1 or 2 femtoseconds. By constraining the fastest bond vibrations which are those involving hydrogen atom bonds, to constant values using specifically designed Lagrange multiplier based algorithms (for example SETTLE, SHAKE and RATTLE), time steps can sometimes be increased up to about 5 femtoseconds.

Molecular dynamics thus proceeds in steps, in which the greatest amount of computational time is spent on calculating the potential energy function. As mentioned, there is a need to simulate time lengths long enough to observe events of interest. This is called the sampling problem in molecular dynamics: when an event of interest is not observed in a simulation one can only speculate whether the event does not happen due to the actual physics and chemistry of the system or if it does happen but we have simply not simulated for a long enough time. The problem is further complicated by inaccuracies in the force fields as well as physical limitations in the accuracy of floating point operations on modern day computers: these small inaccuracies tend to add up as the simulation progresses often leading to instability before the target time scale is reached. The first protein simulation in 1977 was 8.8 picoseconds (ps) long. The longest simulations to date have attained the millisecond time scale, but the great majority of simulations performed today range from 10s to 100s of nanoseconds (ns).

Since the great bulk of calculation time is spent on the potential energy function, a number of approaches aim increase the efficiency of this part of the method. For example, the three bonded term calculation can easily be split up between several computer processers and thus calculated in parallel. Because the Lennard-Jones potential decreases very quickly with distance (*1/r6*), a cut-off can be introduced to only calculate the energy over pairs of atoms that are sufficiently close together. This is a significant savings as the number of pairs of atoms increases as *N2*. A number of sophisticated algorithms, especially parallel computing ones, exist for efficiently maintaining and updating the list of particle pairs within the cut-off distance. The same cannot be done for electrostatics, which decreases much more slowly with distance (*1/r*). Fortunately however, when dealing with a periodic system, an algorithm called Ewald summation (after the same Paul Peter Ewald we met in crystallography) is able to accurately calculate the electrostatic energy by decomposing the interactions into short-range and long-range terms and calculating the Fourier transformed long-range terms in reciprocal space. In 1993 Thomas Darden and Darrin York devised a method called Particle Mesh Ewald (PME) that is able to calculate the Ewald sum in significantly faster time by spreading the charge density on a three dimensional grid.

Finally, we mention that molecules are very rarely simulated in molecular dynamics *in vacuo*. Not only is this unrealistic as molecules usually are not encountered in nature in solitary confinement, but it also would lead to sever artefacts on account of the high energy of electrostatic interaction between charged moieties within proteins and nucleic acids. Thus the standard approach in molecular dynamics is to immerse the molecule of interest in a water box composed usually of several thousand water molecules surrounding the protein. There are many sets of force field parameters for water of which the most popular ones are spc-e, tip3p, tip4p and tip4p-ew. However, even such a system would not be successful as the box of water with protein would itself be located in a vaccum leading all of the waters to fly away from each other into space. Thus an algorithm called periodic boundary conditions (PBC) is used. This essentially consists of replicating the simulated box out infinitely in space in all directions. In other words copies of the box itself are placed all around it so that when a particle flies out of the box on one side, an identical particle flies in on the other side. This not only surrounds the simulated box with virtual matter preventing an exploding artefact but also essentially creates a periodic system, thus enabling us to use the PME algorithm to calculate electrostatics.

### Goals and outline of reasearch

Both macromolecular x-ray crystallography and molecular dynamics have proven themselves to be extremely valuable methods in the biophysical arsenal. However, both methods also suffer from several limitations. It is the overarching idea in the present work that molecular dynamics simulations of crystals can contribute to resolving some of these limitations.

Crystallography: First sources of error and noise. Sometimes so high (low res) that indeterminate. Simulations if reliable could tell us more to help resolve. Second, time and space average… End up with single static view. Myopic because one best rep view of the average. In fact crystals move (dynamic) and heterogenous. Recent efforts by several group aimed at resolving this. Ensemble refinement. Networks stuff from Fraser. Diffuse scattering from … Move and insights about functions. Also insights about crytals. By simulating over multiple copies (space) and ns (time) we can undo the averaging and get time-space resolved glimpse. A more comprehensive view of the crystal. Also, information about crystals: solvent distribution, etc lead to better refinement techniques. Finally, in itself a better set of priors.

MD is good but usually run as solvated box. Good sense but drawback that not comparing against experimental… Validation from crystals. Look directly at experimentals such as structure, fluctuations, electron density and amplitudes… Can check how well it’s doing and modify ff…

We have thus introduced briefly the two methods of biomolecular crystallography and molecular dynamics in a way that should permit the general understanding of the work that follows. We continue with a specific presentation of the research carried out on the various aspects of molecular dynamics of crystals. Part I deals with fundamental questions of the methodology for carrying out molecular dynamics of crystals. It investigates how to set up and carry out crystal simulations as well as how to analyze them given the unique qualities of the produced data. After investigating simulations of a small peptide crystal (Chapter 2 and 3) we move on to examine results when the presented methodology is applied to more relevant crystals of proteins (Chapter 4) and nucleic acids (Chapter 5). Chapters 2 and 3 are also demonstrate the proof of concept that information gained from molecular dynamics of crystals can be directly used to improve our knowledge about crystals and to enhance the interpretation of experimental crystallography data. Chapters 4 and 5 on the other hand examine various aspects of molecular dynamics force field validation and paint possible further paths for improving on current force fields based on data obtained from crystal simulations. Chapter 2 first appeared in the *Journal of the American Chemical Society* as… Chapter 3 first appeared in *Protein Science* as… Chapter 4 first appeared in *Biochimica et Biophysica Acta* as…

Part II presents examples of how molecular dynamics of crystals can play a useful role in the study of various questions in molecular biophysics. In Chapter 6 various states of the hairpin ribozyme are examined, both active precursor and transition states, to shed light on the specific mechanism of the enzyme’s self-cleavage reaction and to propose further paths for experimental validation. Chapter 7 studies Rnase A by a combination of solution state and crystal state molecular dynamics simulations in order to elucidate the mechanism of the phosphoryl transfer reaction it catalyzes. Chapter 6 first appeared in the *Journal of the American Chemical Society* as …

Finally, in Part III we show that molecular dynamics of crystals can be directly employed to improve crystallography methods. In Chapter 8 a molecular dynamics force field is implemented to accurately model protein ligands and small molecules in macromolecular crystals thus leading to chemically more accurate ligand geometries. Chapter 9 presents an integration of the Amber molecular dynamics software package with Phenix software for crystallographic refinement. Incorporation of molecular dynamics of crystals directly in biomolecular crystal refinement leads to improved structural models and better agreement with experimental data. Chapter 8 has been submitted for publication and is currently under review at *Acta Crystallographica D*.

# Developing molecular dynamics of crystals.

## Fav8 1

Blah blah

## Fav8 2

Blah blah

## 4lzt

Blah blah

## DNA/RNA

Blah blah

# Applications of molecular dynamics of crystals

## Hairpin

Blah blah

## RnaseA

Blah blah

# Improved crystallographic methods through crystal molecular dynamics

## AFITT

Blah blah

## Phenix-Amber

Blah blah

# Bibliography

1. Alder BJ, Wainwright TE. Studies in Molecular Dynamics. I. General Method. 1959;31(2):459.

2. Case DA, Babin V, Berryman JT, Betz RM, Cai Q, Cerutti DS, Cheatham, III TE, Darden TA, Duke RE, Gohlke H, et al. AMBER 14. 2014.

3. Hess B, Kutzner C, van der Spoel D, Lindahl E. GROMACS 4:  Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. 2008;4(3):435–447.

4. Hynninen A-P, Crowley MF. New faster CHARMM molecular dynamics engine. 2014;35(5):406–13.

5. McCammon JA, Gelin BR, Karplus M. Dynamics of folded proteins. 1977;267(5612):585–590.

6. Rahman A. Correlations in the Motion of Atoms in Liquid Argon. 1964;136(2A):A405–A411.