Molecular dynamics of sense and sensibility in processing and analysis of data
Wassenaar, Tsjerk

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2006

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
Chapter 1

General Introduction
1 Introduction

“All biochemical systems exhibit dynamic behaviour. In many cases the physiological property itself is in essence a dynamic property, as in enzyme catalysis, active ion transport across membranes, repression and activation of gene expression, and so on.”

_Cantor and Schimmel_  
_Biophysical Chemistry_

1.1 Proteins

A protein is a complex organic molecule consisting of tens to several hundreds of distinct building blocks, called amino acids. A protein can be thought of as a small apparatus dedicated to a specific task, such as the break down of food components and toxic compounds or the synthesis of molecules needed for proper functioning. Though some proteins solely serve as static supports or carriers of other molecules and can be considered inert, for the majority of proteins the function is ultimately linked to the specific mechanic and chemical properties. In other words, for most proteins the function is the result of dynamical processes.

Proteins are vital for the proper function of organisms. In fact, many diseases can be traced back to the malfunction of proteins. The cause of the malfunction can either be intrinsic, i.e. the result of a mechanism in the protein itself failing, or extrinsic, e.g. due to the failure of controlling mechanisms. Examples of intrinsic malfunctions are mutations which alter the mechanical or chemical properties of proteins, usually rendering them less effective or even inactive. Hereditary diseases generally fall in this category. Infections and toxification are classical examples of extrinsic factors causing failure of protein function. Bacteria and viruses often directly counteract proteins which may be harmful to them, causing failure of important processes in the body.

Understanding the normal function and (possible) malfunction of proteins is of vital importance for current day medicine. The study of this is called molecular pathology or molecular pathophysiology. Insight in the processes under normal and pathological conditions (illness) allows the search for therapies aimed at normalizing the functions on the most basic level, the proteins.
1.2 Cells

The cell is the smallest functional unit of life. The human body, for example, consists of an estimated \(10^{14}\) or 100 000 billion cells. The objective of individual cells, notably unicellular organisms, is simple: sustain life. In complex multicellular organisms (metazoans) such as ourselves, this aim is made subordinate to the life of the total organism. In some cases this means that cells have to be sacrificed for the sake of the organism as a whole. In most metazoans the cells possess a special enzymatic pathway solely to effect the self-sacrifice of the cell if the surrounding tissue ceases to provide supporting signals or if the cell receives specific signals to eradicate itself. This process of self-sacrifice is called programmed cell death or, more specifically, apoptosis. Apoptosis will be discussed in more detail below, but first I will illustrate how cells are able to sense that it is still sensible to sustain life or more general, how cells are able to detect signals from their surroundings and act appropriately, within the framework of the organism.

1.3 Sense on the cellular level

Maintaining oneself on the cellular level generally involves constant adaptation. In particular, organisms need to be able to react appropriately to changes in their environment. The reaction can involve an adaptation in their metabolism, moving to where more food is available or away from harm. In order to react to changes, an organism must be able to detect changes in its environment. For this, all cells have specialized proteins on their outer membrane, dedicated to sensing. These are the cell-surface receptors. Stimulation of cell surface receptors, which can be chemical (e.g. chemical compounds, pH) or physical (e.g. light, temperature and pressure), leads to the modulation of intracellular processes and thus to a reaction of the organism to the changes in the surroundings.

In multicellular organisms, down to the most primitive forms, the mechanisms for sensing are not only used to gather information on the environment per se, but also for communication between cells. In this way, the cells can influence the behaviour of others and cooperate. This allows groups of cells to perform specialized tasks and opens the way for the formation of complex organisms, in which groups of cells are highly specialized.

In complex metazoans, such as humans, the degree of differentiation is enormous. Maintaining the internal environment (homeostasis), involves a manifold of different mechanisms even with regards to a single property (e.g. the blood pressure). Usually, each of these mechanisms is governed by specialized cells. Often, as in the case of the blood pressure, the sensor and effector cells have been decoupled. The actual blood pressure regulation is the result of intense communication between the sensor cells and the brain and between the brain and the various effector cells, e.g. located in the heart, the blood vessels and the kidneys.

To allow specific communication, many different cell surface receptors have evolved. In particular receptors for neurotransmitters or hormones can enable the body to alter its metabolism on a cellular level in order to maintain homeostasis.

1.4 Differentiation

A particularly important process in metazoans with a high degree of differentiation is the establishment of this differentiation during development. In addition, the life span of most cells is shorter than that of the organism and cells need to be replaced by equally differentiated cells. The
processes governing differentiation in the earliest stages of development are still poorly understood. However, many differentiation pathways later in the development are well known and involve specific factors or hormones, which often function by activating specific receptors. An example of a differentiation pathway is the development of red blood cells from precursor blood cells (stem cells) in the bone marrow. This process is mediated primarily by erythropoietin, a glycoprotein (a protein linked to sugar groups) which is synthesized in the kidneys and the liver. Erythropoietin functions by binding to a dimeric cell surface receptor, activating the intracellular domains and associated proteins. In turn, this leads to activation of gene transcription modulators. Erythropoietin (EPO) is best known for its misuse in endurance sports, such as cycling and long-distance running. The increase in red blood cells resulting from administration of EPO increases the oxygen uptake capacity, boosting performance. The molecular mechanism of activation of the erythropoietin receptor (EPOR) is the central topic of Chapter 6.

1.5 Apoptosis

As mentioned above, another process of vital importance to metazoans is apoptosis. There are three mechanisms which can lead to activation of a cell’s apoptotic machinery. The first of these involves the response to irreparable damage to the DNA, e.g. by x-rays. Another mechanism is activated if a cell looses the contact with its environment in which case the cell gets deprived of life signals. The third mechanism is the most interesting one, because it involves direct activation of the apoptotic machinery by external factors. Apparently, the availability of escape mechanisms to eradicate redundant or possibly harmful cells is vital to sustain a complex society of differentiated cells such as ourselves.

The external factors which can stimulate a cell to undergo apoptosis are usually proteins known as cytokines. In particular, certain members of the Tumor Necrosis Factor (TNF) superfamily are responsible for the direct activation of the apoptotic pathway. The most important of these are TNF, Fas (CD95) and the TNF-Related Apoptosis-Inducing Ligand (TRAIL). Among these three, TRAIL is of special interest for its ability to induce apoptosis selectively in malignant (cancer) cells, while leaving healthy cells unaffected.

TRAIL exerts its functions by interacting with two receptors, the Death Receptor 4 and the Death Receptor 5, which are bound to the cell surface. Both of these receptors contain an intracellular domain which is in direct contact with the intracellular effectors of apoptosis. An interesting aspect in TRAIL mediated apoptosis is the presence of two decoy receptors, which are present on healthy cells, but not on certain malignant cell lines, and bind TRAIL without inducing apoptosis.

The interaction of TRAIL with Death Receptor 5 is analyzed in Chapter 5.

1.6 Signal transfer across the cell membrane

It is evident that the binding of a ligand to the extracellular domain(s) of a cell surface receptor leads to the activation of the intracellular domain. However, the mechanisms underlying these processes are less well understood. Essentially, there are two different types of receptor. First, there is the ionotrophic receptor, which functions as an operated ion channel. Activation by a physical stimulus or a ligand will cause the channel to open and change the concentration of a certain species of ion, directly affecting the function of many proteins.
Next to the ionotropic receptors there are metabotropic receptors. These can have a catalytic intracellular domain, the activity of which is mediated by the stimulus, or they are connected to a second messenger system, which is activated after the receptor is stimulated. The metabotropic receptors can be subdivided based on the mechanism by which they act. For example, some multimeric receptors require cross-linking and alignment of the intracellular domains in order to be active. Alternatively, a stimulus can lead to a conformational or allosteric change, changing the characteristics of the intracellular side and affecting the interaction with a second messenger system.

In all of these cases, the opening of an ion channel, the activation by alignment of intracellular domains and the induction of a conformational change, the molecular mechanisms are only poorly understood. However, knowledge of these processes is both vital for understanding receptor function as well as for understanding how to modulate receptor function. An investigation of the possible mechanisms by which TRAIL may induce transmembrane signalling of Death Receptor 5 is also presented in Chapter 5. An investigation of the possible mechanisms of signalling by the EPO receptor also forms part of Chapter 6.

1.7 Cell-surface receptors and evidence-based drug design

It is easy to imagine that the malfunction of cell-surface receptors, either due to intrinsic factors (mutations) or extrinsic factors such as insufficient or excessive synthesis of stimulant, can lead to severe problems in maintaining homeostasis, or rather, to disease. On the other hand, these mechanisms offer opportunities for therapeutic intervention on the cellular level, stimulating or inhibiting groups of cells. For these reasons, the cell-surface receptors are the most important class of pharmacological targets.

The usual practice in medicinal chemistry to generate compounds with the desired activity for a specific receptor is to compare a range of known compounds with regard to the affinity and activity in relation to the structural properties. For example, three-dimensional quantitative structure-activity relationships (3D-QSAR) are commonly used to develop and optimize new potential drugs. Nowadays, these techniques are generally combined with structural models of the targets themselves to improve the fit of the QSAR model. Notably, docking of ligands to proteins and estimation of induced fit effects has greatly improved the efficiency of rational drug-design. However, in this process the dynamic properties of the receptor are frequently ignored. The targeting molecule must either induce or inhibit a certain process in the receptor and knowledge of these processes can be used in evidence-based drug design.

In Chapter 6, the interactions between a set of active and inactive peptides and the EPO receptor are compared to investigate the discriminants of activity and inactivity in terms of the conformation and dynamics of this receptor. In general, the work in this thesis is intended as a step in the use of information from the dynamics of proteins in rational drug design. For example, in Chapter 2 several methods are presented which can be used to investigate the effect of ligands on protein dynamics. In addition, Chapter 4 presents two assessments of the similarity of sets of simulations performed under different conditions. These methods are also applicable to compare e.g. the effect of different ligands on a protein.
2 Molecular Dynamics

Unfortunately it is difficult to obtain structural information on membrane proteins such as cell-surface receptors. Though advances have been made, the number of high-resolution structures of cell-surface receptors available in the Protein Data Bank[1] is still limited. Moreover, none of the currently available experimental techniques allows one to examine the mechanism of activation of these proteins in atomic detail.

If the structure of a protein is available, the dynamics and consequently the mechanism of action can be probed using computational methods, notably molecular dynamics simulation techniques, which allow one to investigate intra- and intermolecular processes at atomic resolution.

2.1 Particles and Forces

All phenomena in nature can in principle be understood in terms of particles and forces acting on particles. There are four fundamental forces known at present. The first is the electromagnetic force, which acts on charged particles. The second and third are the weak nuclear force, which keeps the protons and neutrons together in atomic nuclei, and the strong nuclear force, which keeps the quarks inside protons and neutrons tightly bound to each other. The fourth fundamental force is the gravitational force, which keeps planets in orbit around stars and moons in orbit around planets.

2.2 Proteins with no strings attached

A protein consists of atoms, which are built of smaller components, notably protons, neutrons and electrons. But to understand the dynamics of a protein, it is not necessary to regard the processes deep inside atoms. Nuclear fission and fusion are not biologically relevant processes and it is safe to disregard the two nuclear forces.

Likewise, the importance of gravity is negligible for the function of proteins on the molecular level, although it is an important force in daily life on the scale of humans. Neglecting the contributions of gravity leaves one important force, which is relevant for the function of proteins: the electromagnetic force. Together with the particles on which it acts, i.e. nuclei and electrons, this force forms the basis for computational methods in protein science.

If the desire is to investigate the mechanical behaviour, rather than chemical processes, it is possible to make a further simplification. The quantum mechanical effects and electronic degrees of freedom can be ignored or effectively treated as constraints. A protein is then effectively regarded as a mechanical construct, consisting of soft spheres with a certain connectivity and a set of empirical functions can be used to describe the interactions between particles. This is the classical approximation.

In the classical approximation, an atomistic system is deterministic. If the positions and momenta of the atoms and the forces acting upon them are known, it is possible to calculate both the past and the future. Essentially, this is the world of Laplace's daemon:
“We may regard the present state of the universe as the effect of its past and the cause of its future. An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, it would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom; for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes.”

Pierre Simon de Laplace (1749-1827)

This is the basis of classical molecular dynamics simulations.

### 2.3 Molecular Dynamics

The molecular dynamics simulation technique is often attributed to the pioneering work of Alder and Wainwright in the late 1950s[2]. However, the term molecular dynamics was already coined at the end of the 19th century by William Thomson, better known as Lord Kelvin, who also gave the first known account of molecular dynamics calculations in his book, the Baltimore Lectures on Molecular Dynamics and the Wave Theory of Light[3]:

“I desire to take this opportunity of expressing my obligations to Mr. William Anderson, my secretary and assistant, for the mathematical tact and skill, the accuracy of geometrical drawing, and the unfailingly faithful perseverance in the long-continued and varied series of drawings and algebraic and arithmetical calculations, explained in the following pages. The whole of this work, involving the determination of results due to more than five thousand individual impacts, has been performed by Mr. Anderson.”

William Thomson (Lord Kelvin), 1901

The basic principle behind molecular dynamics is the second law of Newton, which states that “an applied force is equal to the rate of change of momentum” or the force applied to a body is equal to the mass of the body times the acceleration:

\[
F = \frac{dp}{dt} = ma
\]

(1.1)

Given that the mass of each particle in the system is known and the forces can be derived from the interactions with the surrounding particles, the acceleration of each particle can be calculated. From this the instantaneous velocity and the displacement can be calculated according to

\[
\frac{dv_i(t)}{dt} = \frac{F(t)}{m_i}
\]

(1.2)

and

\[
\frac{dr_i(t)}{dt} = v_i(t)
\]

(1.3)
The force $F_i$ exerted on atom $i$ by the other atoms in the system is given by the negative gradient of a potential energy function $V$, which in turn depends on the coordinates of all other atoms in the system:

$$F_i(t) = -\frac{\partial V(r_i(t), r_2(t), \ldots, r_N(t))}{\partial r_i(t)}$$ (1.4)

The potential energy function is based on a set of interaction functions and parameters, called the force field. Both the functional form and the parameters of the force field are in general derived empirically. Between different force fields, differences can exist both in the functional forms and the parameters, though the force fields used for molecular dynamics applied to biomolecules, such as proteins, have similar general forms. In all cases, the interactions fall into two categories: bonded and non-bonded interactions. Usually any two atoms which are connected by one, two or three bonds are treated through bonded interaction terms, whereas interactions between more distant atoms and between atoms which are not connected at all are described using non-bonded terms. In most cases there are two types of non-bonded interactions, namely coulombic interactions between charged particles and dispersion interactions generally described by a Lennard-Jones potential. The bonded interactions generally include bonds (1–2 interactions), angles (1–3 interactions) and torsion angles (1–4 interactions). As an example, the effective potential for the GROMOS force field[4-6] can be written as:

$$V(r_1, r_2, \ldots, r_N) = \sum_{\text{bonds}} \frac{1}{2} k_b (b^2 - b_0^2)^2 + \sum_{\text{angles}} \frac{1}{2} k_{\theta} (\cos \theta - \cos \theta_0)^2 + \sum_{\text{dihedrals}} k_{\phi} (1 + \cos(\phi)) \cos(m) + \sum_{\text{improvers}} \frac{1}{2} k_\beta (\xi - \xi_0)^2 + \sum_{\text{coulomb}} \frac{q_i q_j}{\varepsilon r_{ij}} + \sum_{\text{LJ}} \left( \frac{C_{ij}^{(12)}}{r_{ij}^{12}} - \frac{C_{ij}^{(6)}}{r_{ij}^{6}} \right)$$ (1.5)

In this function, $k_b$, $k_{\theta}$, $k_{\phi}$ and $k_\beta$ denote the force constants for the bonds, angles, dihedrals and improper dihedrals, respectively.

The core of any MD algorithm is the calculation of the potential on each atom in the system and the subsequent calculation of the displacement over a very small time step, typically in the order of femtoseconds ($10^{-15}$ s). This procedure is repeated for many steps, typically millions, to obtain a trajectory, which contains the positions of all atoms sampled at regular intervals.

Though the calculation of the displacements is essentially as given by equations (2) and (3), the actual implementation of the integration can be different. For example, a commonly used scheme is the leap-frog algorithm used in Gromacs[7-9]:

$$\mathbf{v}_i(t + \Delta t/2) = \mathbf{v}_i(t - \Delta t/2) + \frac{\mathbf{F}_i(t)}{m_i} \Delta t$$ (1.6)$$

$$\mathbf{r}_i(t + \Delta t) = \mathbf{r}_i(t) + \mathbf{v}_i(t + \Delta t/2) \Delta t$$ (1.7)
As can be imagined, molecular dynamics simulations are computationally very demanding. This is particularly true when dealing with large systems such as receptor – ligand complexes. One way to increase the efficiency is to reduce the volume of the unit cell. A method for this is presented in Chapter 3.

2.4 Analysis

The data obtained from a molecular simulation are a series of successive configurations of the atoms as a function of time. The analysis of results obtained from molecular dynamics simulations usually involves the derivation of system specific properties from this series of configurations. These properties, such as the solvent exposed surface and the secondary structure content of a protein, depend on the instantaneous configuration and for this reason they will be referred to as instantaneous properties. This stands in contrast with properties which depend solely on the constitution of the system, such as the number of amino acids, the number of atoms and the number of charged species, which (in simulation) are independent of time and are therefore referred to as characteristic or intrinsic properties. Besides these there are properties which can only be defined over a given window of time, e.g. the viscosity. These are commonly referred to as dynamic properties.

The instantaneous properties are the primary observables for the characterization of a simulation and the comparison of simulations to experiments. These properties can also be used to compare the results from two or more simulations. Note that these properties are often determined as an average value over a certain time window.

To properly compare simulations based on a set of properties derived from these simulations, it is necessary to understand the nature of these properties and, more importantly, the variation to be expected between otherwise equal simulations.

To assess the similarity or otherwise of sets of simulations one must estimate the “natural” variation and take this variation into account during qualitative analysis. For this it is necessary to use statistical tests, tailored to molecular dynamics data. A theoretical framework for such tests is presented in Chapter 2, and its application to assess the effect of external or applied conditions on the outcomes of simulations is presented in Chapter 4.

2.5 Collective motions

The connectivity of atoms in a protein, through bonds, angles and dihedrals, as well as the non-bonded interactions between neighbouring atoms gives rise to correlations between the motions of atoms. For this reason, another approach to the analysis of molecular dynamics data is the investigation of the concerted motions of atoms, which are generally thought essential for the function of the protein. This type of analysis finds its basis in principal component analysis, which is a statistical method for data reduction. Applied to the coordinates obtained from a simulation it reveals any structure in the atomic fluctuations, and allows insight in the internal mechanisms of action.

Principal component analysis in molecular dynamics comes in different flavours. All of these methods are however aimed at understanding the collective motions of a single system. Because many problems in biology and pharmacy are concerned with multicomponent systems, such as protein – ligand and protein – protein interactions, rather than individual proteins, it is desirable to investigate the system in such a way that information is given in terms of single subsystems and interactions between subsystems. Various methods for this type of analysis are presented in Chapter 2.
In addition, in Chapter 2 several methods are presented for the investigation of the relationship between collective motions and the simulation conditions applied or observed instantaneous properties. Most of these methods were derived during the work performed for this thesis, to assess specific questions.

3 Objectives of the project

The objectives of the project are:

1. Design and application of statistical methods for processing and analysis of data obtained from molecular dynamics simulation, with specific attention for protein–protein and protein–ligand interactions.
2. Design of methods to increase the efficiency of simulations, notably by reduction of the size of the simulation system.
3. Application of such methods to gain insight in the functional mechanisms of large cytokine receptor systems, namely the TRAIL receptor Death Receptor 5 (DR5) and the Erythropoietin receptor (EPOR)
4. Application of such methods to gain insight in the function of bacteriorhodopsin, with emphasis on understanding the role of calcium ions in that system.

4 Outline of the thesis

The thesis consists of eight chapters, including this general introduction and an introduction to methods for statistical analysis of data obtained from molecular dynamics simulations, work associated with the Near-Densest Lattice Packing, the TRAIL – DR5 complex, the EPO – EPOR complex and the Purple Membrane. The final chapter contains some overall conclusions, perspectives and a summary. A more detailed description of each chapter is given below:

In Chapter 2 some elementary concepts of (multivariate) statistics are explained, together with applications for the analysis of molecular dynamics simulations. This chapter will combine different methods of analysis used in the subsequent chapters and place them in a common statistical framework. In addition, a number of methods are introduced, which are not directly used for the results presented in this thesis, but were developed in the course of the work to deal with specific problems.

Chapter 3 discusses a method to construct a simulation box based on the geometry of the solute under study. This chapter is complemented with appendices in which further aspects concerning simulation cells and the determination of nearest images are considered.

Chapter 4 presents two studies involving a statistical comparison between simulations. The first of these studies is concerned with the effect of the box type on the simulation outcome. The second study assesses the effect of a change in the force field on simulations performed on 31 different proteins.
In Chapter 5 the interaction between the cytokine TRAIL and its receptor, Death Receptor 5 (DR5), is discussed. This chapter provides a possible mechanism of activation of the receptor. In addition, the formation of functional supramolecular structures and a possible mechanism for reverse signalling by TRAIL are discussed.

Chapter 6 deals with the interaction between EPO and its receptor EPO-R, where the difference between the ligand bound and unbound receptor is discussed, together with the asymmetric binding of EPO. In addition, an assessment is made of determinants of binding affinity and activity in the characteristic motions of EPOR, in conjunction with a set of EPO mimetic peptides.

In Chapter 7 a model for the Purple Membrane is presented, and the results obtained on the binding of calcium to the purple membrane are discussed.

5 References
