Chapter 1

Introduction
This shift towards developing complex models that can describe complex systems can already be observed in a number of fields of research, like economics, various engineering practices, meteorology, quantum chemistry and protein folding, in which apparently the fruitful approaches involve large-scale simulations of the system in question. Often the level of simplification is dictated by the available computing power (e.g. in meteorology), or, conversely the maximum level of simplification required for a certain level of reliability sets a limit on the size of the systems that can successfully be taken under consideration (e.g. in quantum chemistry).

Continuing along this line of thought, it seems that major breakthroughs in years to come will most likely be linked to sophisticated and clever methods for handling and analyzing huge datasets of a very complex nature. A broadly used but still relatively simple example of this is the mathematical method of principal components analysis, which can extract general trends and correlations from a (large) set of variables. It also appears likely that cross-overs from one field of research to another will be more frequent and will have high chances of success, since these types of methods will probably have more or less general applicability, and will therefore not be restricted to a particular field. In other words, we should see an increasing emphasis on mathematical methods and models specifically equipped to deal with these large complex data sets.

1.2 The Organismic View

There is one area of research that might be described as being “meta-complex” as it involves aspects of several of the complex areas listed above. It can be captured under the broad label of “life sciences”, involving the interaction between the molecular basis of life (quantum-physics and biochemistry), its evolutionary history (molecular biology and biology), the environment and eco-system (ecology and sociology) and its possible applications in engineering (nano-technology). This makes it a very challenging field of research, and considering our own interest as H. sapiens the amount of interest it receives should not be a surprise.

Much discussion is possible about which of the complex systems associated with living organisms is most fundamental to our existence, or that of terrestrial organisms in general. Mostly, it will be a matter of taste. However, there appears to be one particular thin spot in our general understanding of life. From one side, we know how the process of evolution causes the prevalence of certain genes, and we know how genes give rise to proteins. From the other side, we know how organisms are organized and how they are “constructed” out of organs, and how those are constructed out of cells. We also know what a cell is made of, and in which way many of its processes are regulated. Many of these areas have their own problems in understanding certain aspects, but there is one thing that ties everything together of which the basic mechanism remains unknown. While we know how the genetic code is translated into a sequence of amino acid residues, and we know of many proteins how they function, the understanding of how the sequence translates itself into the functional folded protein remains elusive.
1.1 The Cosmic View

The whole of contemporary science appears to be headed towards a collective complexity barrier, in which the wide variety of systems under study share one common characteristic of being of such complexity, that a conventional description of well-defined constituent parts with well-defined interactions, no longer seems able to capture the nature of the phenomenon under observation. Still worse, in some areas even the distinction between observer and observed cannot clearly be made. Areas in which these problems appear particularly obvious, are:

- **quantum-physics**, where an “elementary particle” is never isolated but always accompanied by a shroud of “virtual particles”, and interactions are thought to be transmitted by yet more particles (worse still, the distinction between virtual and real particles, and between particles and interactions, is actually meaningless in that context);

- **nano-technology** where matter must sometimes be described as collections of atoms, or as quantum waves, or in terms of averages and densities and often new, hybrid and still poorly understood models are necessary to describe properties that emerge from the interplay of these descriptions;

- **biochemistry**, where the whole cell is seen to be a continually changing and cycling machinery of molecules, of which probably most have an influence on each other;

- **biology or ecology**, where it has become clear that flora, fauna, soil, water and air are inextricably linked into an eco-system that we once thought could be divided into independent parts, and we are gradually learning (the hard way) that we ourselves are an equally inextricably linked part of it;

- **sociology** and **economics** (which can be viewed as a branch of sociology with a different set of boundary conditions), in which often even the act of performing the experiment, and afterwards publishing the result, cannot be separated from the experiment itself;

- **cosmology**, which is linked to particle and quantum-physics, bringing the circle back round to where we began.

This might mean that the historical approach of science of bringing down the complexity of the world around us into a limited set of simple, well-defined rules, is coming to an end and that the remaining big questions can only be answered by tackling them in all their immense complexity and detail. Fortunately, we have recently also begun developing tools which are very well equipped for handling the enormous tasks of analyzing systems of high complexity in a highly detailed fashion. The most obvious of these is the computer which allows the storage, retrieval and processing of the raw data associated with such systems. Equally important are the mathematical methods and models that can minimize the complexity of the raw data describing the system and thus uncover the intrinsic complexity of the system itself. The compression algorithms used for image and audio data are a nice example.
1.3 The Molecular View

All beings (as we know them) consist of molecules. The most interesting ones are the so-called “biological macromolecules”, namely proteins (which make up the cell’s machinery), DNA and RNA (which respectively store and transport the genetic information) and lipids (which compartmentalize the processes in the cell, and the cell itself from the outside). Proteins are the molecules in a cell that accelerate and regulate all processes involved in making a living cell alive. That, in a sense, makes them central to the understanding of how a cell functions. At the same time, some of the key features of protein molecules are poorly understood.

What is true for the life sciences as a whole, also holds for proteins. For most of the research areas involving highly complex systems, as mentioned above, one can identify area’s of scientific research directed towards understanding of the working of proteins, or to the relations of a protein with its surroundings. The function of a protein as an enzyme, that accelerates a specific chemical process, can only be described and understood using a proper quantum-chemical description. The existence of a certain protein with a certain appearance and construction must be described in its evolutionary context, which brings it also in relation to its present and past functions in the cell and via that to possible roles that might have in the broader context of the organism and society. The specific shape and function a protein has is inextricably linked to its construction, which is derived from the descriptions in the genetic code, which can only be understood, again in the evolutionary context and its function in the cell, but also against the background of the basic physical knowledge of the behavior of individual atoms and molecules. The behavior of large protein complexes, like the keratin fibers that hairs are made of, or muscle fibers which are made out of countless units of actin and myosin, or the aggregates of proteins that form the amyloid plaques that can ravage a person’s (or animals) brains, might only be comprehensibly described on a mesoscopic level, beyond the level of individual molecules.

Previously, the so-called “protein-folding problem” was already briefly mentioned: how does a sequence of chemically linked amino acids, in total several thousands of atoms, fold itself into a specific structure? This structure often is defined to within a tenth of a nanometer, much less than half the diameter of the typical atom in proteins, a carbon atom. This cannot happen purely by chance, if one works out the odds it is impossible and this is known as the “Levinthal paradox”; protein folding appears impossible but yet it happens. Therefore some mechanism must govern the process that leads from an unstructured, freshly produced sequence of amino acids, to this fully folded, functional, ‘native’ protein.

Broadly speaking, three approaches are being pursued in the field of protein folding research: biochemical, structural and computational. In the first, one attempts to learn all there is to learn in a laboratory about the functions that can be performed by a protein, modified versions of the protein, and more or less related ‘cousins’ of the protein from different species. The second approach, also experimental, elucidates the appearance of a protein by determining the three dimensional position of all its atoms, from which an extremely detailed ‘picture’ of the protein emerges, known as the “protein structure”. The third approach is theoretical and takes a known three dimensional structure of a protein,
the known properties of its atoms (like radius, mass, charge and connections or bonds),
Newtons’ laws of motion and a powerful computer to predict the possible motions of all the
atoms that could occur within the protein, i.e. the dynamics of a molecule are simulated.

One important point has been made clear in the last couple of years, however, that
neither simulations nor experiments alone will be able to solve the issues at hand. Gener-
ally speaking, experiments lack sufficient detail and discriminatory power at the atomic
level and at short timescales, and conversely, simulations lack sufficient statistics on mul-
tiple (or large) molecules and at long timescales. The way forward therefore must lie in a
fruitful synthesis of simulation and experiment.

There is a two-fold hurdle in combining simulation and experiment derived data for
biological molecular systems. This lies in the difference in timescale accessible to simu-
lation and experiment, and in the comparison of the raw data of both. The timescale
gap can be bridged by (ever increasing) more sophisticated experimental set-ups and by
performing simulations by more efficient means on, also ever increasing, faster computers.
Comparison of data can best be done by calculating from the simulations the actual raw
experimental data as directly as possible.

1.4 The Local View

The hypothesis that started off the work described in this thesis, was about an aspect
of protein folding. Early engineered four helix bundle “proteins” lacked the well-defined
interior structure characteristic of “natural proteins”, instead giving rise to a state often
referred to as a “molten globule”, indicating the presence of secondary structure (like he-
lies) and the absence of a packed structure for the amino acid side chains in the interior.
Introduction of a covalently bound heme group in the interior of the helix bundles, yielded
an anchoring point around which a “natural” protein structure could form. From these
results, the thought arose that also in naturally occurring proteins such an anchoring point
must be present, which would induce structure in the native state structure. This anchor
could consist of a salt-bridge, a sulfur bridge, a buried hydrogen bond, a stacking inter-
action between aromatic side chains, or an interdigitation between a number of aliphatic
side chains. Due to this diversity in the nature of possible anchoring points, it seemed
not to be feasible to identify them directly by searching for the constituent components.
Further thinking centered around ideas of possible modes of motion of the protein around
such an anchoring point, in terms of single- or double-hinged, swiveling or sliding motions.
Identifying these types of motion in the protein dynamics could be a way of finding the
anchoring points.

One approach for identifying anchoring points was based on the influence they should
have on the correlation of motions of the surrounding protein. Therefore, a thorough
analysis was performed of motional correlations in a protein in its native state as expressed
during molecular dynamics simulations. An anchor would reveal itself by correlation of
spatially adjacent but sequentially distant parts of the protein. Instead it was found that
all significant correlation in a protein is sequential. This, added to other negative results
from the search for anchors, effectively negated the initial hypothesis. It was replaced by
1.4. The Local View

the notion of “motionally coherent elements” that form the elementary building blocks of a protein in a dynamical sense. Some of the preliminary work was written up in a conference proceedings as “The Domain Decomposition of a Single-Domain Protein” [7]. The results and reasonings that led to these (and other) conclusions are presented in chapter 3. The findings make it possible to generate a simplified description of the dynamic nature of a protein.

In the context of protein folding, dynamics obviously plays a significant role. Therefore, the “motionally coherent elements” mentioned above, might in addition be related to dynamical elements that behave more or less independently during folding, thus they might be referred to as “folding units”. Some of the possible implications of this hypothesis are investigated in chapter 5. The elements identified from the protein are treated as isolated peptides in solution. A tendency of such a peptide to adopt the protein conformation can indicate a possible role of the element in the early stages of the protein folding process. In chapter 6 one of these fragments that showed most promising as possible folding nucleation site, is examined in detail using NMR experiments and MD simulations.

A large portion of this thesis is devoted to making steps towards overcoming the hurdle of combining simulation and experiment. In simulations, time scales of 1 µs are considered huge, whereas for many experiments such times lie well below the lower limit of temporal resolution or dead-time. Likewise, simulations of ensembles of protein molecules are rare (to say the least), but any experiment will yield results on ensembles of at least $10^{10}$ molecules. Fortunately, recent experimental developments are heading towards the single molecule level. The most spectacular is pulling atomic force “microscopy” (AFM), where a single molecule is caught with the tip of an AFM and subsequently pulled apart. But also single molecule spectroscopic measurements are being developed, and single molecule imaging can be done using electron microscopes. Still, in general these techniques yield results on timescales that are far beyond the accessible time scales of MD simulations, so they do not yet enable direct comparisons. On the side of comparison, the main bottleneck is in the interpretations and assumptions that must often be made while processing experimental data. In chapter 3 aspects of improving simulation efficiency are addressed which extend the time scales accessible in the simulation. In chapter 6 comparisons are made between NMR data and extensive simulations of a nine-residue peptide. The calculated NMR observables will also be used to assess possible artifacts resulting from the different simulation set-ups devised in chapter 3 for increased efficiency.

A different example of dynamical behavior of complex biological molecules and of comparing results from experiments and simulations, is presented in chapter 7 where the fluorescent co-factor flavin adenine dinucleotide (FAD) is studied. The fluorescent behavior of FAD is directly linked to the dynamical behavior, which makes the comparison of MD simulations with fluorescence experiments interesting. In addition, FAD fluorescence measurements can be used as a tool for studying protein to which it is bound, but in many cases also traces of free FAD will be present. For a proper understanding of these experiments, a detailed understanding of the dynamical and fluorescent properties of FAD is important.