DISSERTATION

submitted to the
Combined Faculties for the Natural Sciences and for Mathematics
of the
Ruperto-Carola University of Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

Put forward by
Diplom-Physiker Emal M. Alekozai,
born in Kabul, Afghanistan.

Oral examination: 2nd July 2013

Enhanced Multiscale Sampling of the Cel7A-Cellulose Interaction

A Modelling and Simulation Study to Understand the Enzymatic Conversion of Waste Cellulose into Biofuels

Advisors: Prof. Dr. Jeremy C. Smith
Prof. Dr. Roland Eils

Abstract

Cellulose, the most abundant biopolymer on earth (≈100 billion dry tons/year), holds enormous potential as a renewable energy source. It is a complex carbohydrate that forms the cell walls of plants and gives them rigidity. The cellulose sugar subunits can be unlocked and fermented to produce bioethanol. Plants have developed over time defense mechanism which locks up the sugars and makes the fermentation process difficult. The cellulase enzyme Cel7A is capable to break up this sugar chains, it consists of a carbohydrate-binding module (CBM) and a catalytic domain (CD), joined by a linker peptide. I preformed extensive sets of all-atom Brownian dynamics (BD) ($> 54,600$ trajectories, > 76 ms) and all-atom molecular dynamics (MD) (99 trajectories, $> 6.16\,\mu s$) simulations to study the role of the CBM domain and the linker peptide in the interaction of Cel7A with the cellulose Iβ fiber model. With present supercomputers it is still challenging to study the Cel7A-cellulose interaction on the millisecond timescale at the atomic level. In this work, first an enhanced multiscale framework is derived to combine BD and MD simulations for the Cel7A-cellulose system. Second, it is applied to get new insights in the role of the CBM and the linker peptide.

Cellulose hydrolysis is limited by the accessibility of Cel7A to crystalline substrates, which is perceived to be primarily mediated by the CBM. Therefore, understanding the molecular-level details of the CBM-cellulose fiber interactions are of particular relevance. Here, the binding of CBM to the cellulose Iβ fiber is characterized by combined BD and MD simulations. Coarse-grained BD simulations are used to characterize the diffusional encounter of CBM with different cellulose fiber surfaces, and the site-specific binding results from the BD simulations are then refined via MD simulations to investigate the detailed molecular interactions. The results confirm that CBM prefers to dock to the hydrophobic than to the hydrophilic fiber faces. Both electrostatic (ES) and van der Waals (VDW) interactions are required for achieving the observed binding preference to the hydrophobic fiber faces. At short separation distances, the VDW interactions play a more important role in stabilizing the CBM-fiber binding, whereas the ES interactions contribute through the formation of a number of hydrogen bonds between the CBM and the fiber. At long distances, an ES steering effect is also observed that tends to align the CBM in an antiparallel manner relative to the fiber axis. Furthermore, the MD results reveal hindered diffusion of the CBM on all fiber surfaces, with the diffusion being more restricted on the hydrophobic than on the hydrophilic surfaces. The binding of the CBM to the hydrophobic surfaces is found to involve partial dewetting at the CBM-fiber interface coupled with local structural arrangements of the protein. The present simulation results complement and rationalize a large body of previous work on the CBM binding and provide detailed insights into the mechanism of the CBM-cellulose fiber interactions.

Experiments indicate that the linker peptide might plays a critical role in the cooperative interaction of the CD and CBM with the cellulose fiber. Here, the role of the linker in the Cel7A-fiber encounter process is studied using extensive multiple BD and MD simulations. The linker is represented in the BD simulations as a Hookean spring, with varying length and stiffness. The MD simulations show that the linker peptide properties, including the equilibrium length and spring stiffness, are not intrinsic to the linker but depend strongly on whether explicit or implicit solvent is used, if a fiber surface is present or not, on the hydrophobicity of the fiber surface and whether the complete Cel7A or only the linker is studied. The results further show that the linker has two different states: "extended" and "compact". The BD simulations show that the linker length and stiffness have significant effects on the thermodynamic and kinetic preference of Cel7A for the hydrophobic fiber face of cellulose Iβ, and the mobility of the CBM and the CD-fiber interaction. I further propose a linker length and stiffness optimized for both the CBM-fiber and CD-fiber interactions.

Zusammenfassung

Zellulose ist ein auf der Erde im Überfluss vorhandenes Biopolymer (ca. 100 Mrd. Tonnen Trockengewicht pro Jahr) und hat ein enormes Potenzial als eine erneuerbare Energiequelle. Es ist ein Polysaccharid, welches am Aufbau der pflanzlichen Zellwand beteiligt ist und somit dieser Zelle Stabilität verleiht. Die Zuckereinheiten, aus denen Zellulose besteht, können herausgelöst werden, um zu Bioethanol fermentiert zu werden. Pflanzen haben über die Zeit verschiedene Verteidigungsmechanismen entwickelt welches das herauslösen der Zuckermoleküle und den Fermentierungsprozess erschwert. Das Cellulase Enzyme Cel7A ist in der Lage die Zuckerketten zu zerlegen. Es besteht aus einem Kohlenhydratbindungsmodul (CBM) und einer katalytischen Domain (CD), welche über ein Linkerpeptid verbunden sind. In dieser Arbeit wird eine umfangreiche Reihe von Simulationen - beide in atomarer Genauigkeit - durchgeführt: (I) Brownsche-Dynamik (BD) ($>$ 54.600 Trajektorien, $>$ 76 ms) und (II) Molekular-Dynamik (MD) (99 Trajektorien, $> 6,16\,\mu s$). Ziel ist es die Rolle der CBM Domain und des Linkerpeptides bei der Wechselwirkung von Cel7A mit dem Zellulose Iβ Fasermodell zu untersuchen. Mit heutigen Supercomputern ist es immer noch eine Herausforderung, die Wechselwirkung von Cel7A mit der Zellulosefaser auf der Millisekunden Zeitskala, in atomarer Auflösung, zu untersuchen. In dieser Arbeit wurde ein verbesserter Multiskalenansatz für das Cel7A-Zellulose-System entwickelt, um BD und MD-Simulationen miteinander zu kombinieren. Im nächsten Schritt wurde er angewendet, um neue Einblicke in die Rolle des Cel7A CBM und des Linkerpeptides zu erhalten.

Die enzymatische Zersetzungsrate von Zellulose ist limitiert durch die Zugänglichkeit von Cel7A zum kristallinen Zellulosesubstrat, es wird vermutet, dass sie hauptsächlich durch das CBM beeinflusst wird. Deswegen ist das detaillierte Verständnis der CBM-Zellulosefaser Wechselwirkung auf der molekularen Ebene von entscheidender Bedeutung. Hier wird die Bindung des CBM mit der Zellulosefaser durch kombinierte BD und MD Simulationen charakterisiert. BD Simulationen werden benutzt um den Diffusionscharakter des Zusammenstoßes des CBM mit den verschiedenen Zellulosefaser Oberflächen zu untersuchen. Die Oberflächen abhängigen Bindungsergebnisse der BD-Simulationen werden dann mit MD Simulationen verfeinert, um die detaillierte molekulare Wechselwirkung detailierter zu untersuchen. Die Ergebnisse bestätigen, dass das CBM beim Dockingvorgang die hydrophobe Faseroberfläche über die hydrophile bevorzugt. Die elektrostatische (ES) sowie die van der Waals Wechselwirkung (VDW) sind beide notwendig, um die beobachtete Bindungspräferenz an die hydrophobe Zelluloseoberfläche zu erzielen. Bei kurzen Abständen spielt die VDW Wechselwirkung die wichtigere Rolle beim Stabilisieren der CBM-Faser Bindung. Aber auch die ES Wechselwirkung, trägt durch die Bildung von Wasserstoffbrückenbindungen zwischen dem CBM und der Faser bei. Bei größeren Abständen kann zusätzlich ein ES Ausrichtungseffekt beobachtet werden, der zu einer antiparallelen Ausrichtung des CBM relativ zur Faserachse führt. Zudem offenbaren die MD Ergebnisse eine eingeschränkte Diffusion des CBM auf allen Faseroberflächen, die Diffusion ist stärker eingeschränkt auf der hydrophoben als auf der hydrophilen

Oberfläche. Bei der Bindung des CBM zur hydrophoben Oberfläche wird eine teilweise Wasser-Entnetzung des CBM-Faser Grenzbereiches beobachtet, gekoppelt mit einer lokalen strukturellen Umordnung des Proteins. Die präsentierten Simulationsergebnisse ergänzen und rationalisieren eine große Anzahl bisheriger CBM Bindungsstudien und liefern detaillierte Erkenntnisse zum Mechanismus der CBM-Faser Wechselwirkung.

Experimente deuten darauf hin, dass das Linkerpeptide eine entscheidende Rolle bei der kooperativen Wechselwirkung von CD und CBM mit der Zellulosefaser spielt. Die Rolle des Linkerpeptides beim Dockingvorgang mit der Zellulosefaser wird mithilfe einer umfangreichen Anzahl von BD und MD Simulationen untersucht. Der Linker wird in den BD-Simulationen durch eine Hookesche Feder mit unterschiedlichen Längen und Federkonstanten modelliert. Die MD Ergebnisse zeigen, dass die physikalischen Eigenschaften des Linker (z.B. die Gleichgewichtslänge und die Federkonstante) keine intrinsisch determinierten Eigenschaft des Linkerpeptides sind, sondern dass sie davon abhängen, ob bei der Simulation ein explizites oder implizites Wassermodell benutzt wird. Zustzlich sind die Linker Eigenschaften vom Hydrophobizitätsgrad der Zelluloseoberfläche und ob das gesamtes Cel7A Enzyme oder nur das Linkerpeptid untersucht wird. Die Ergebnisse zeigen weiter, dass das Linkerpeptid zwei verschiedene Zustände hat: “gestreckt” und “kompakt”. Die BD-Simulationen zeigen, dass sowohl die Länge, als auch die Federkonstante des Linkerpeptides einen signifikanten Einfluss auf die thermodynamische und kinetische Vorliebe von Cel7A für die hydrophobe Faseroberfläche von Zellulose Iβ hat. Die Linkereigenschaften beeinflussen ebenfalls die Mobilität des CBM und die CD-Faser Wechselwirkung. In dieser Arbeit wird zusätzlich eine Länge und Federkonstante für das Linkerpeptid vorgeschlagen, die optimiert ist sowohl für die CBM-Faser als auch die CD-Faser Wechselwirkung.

> Knowledge is in the end based on acknowledgement.
>
> *Ludwig Wittgenstein 1889-1951*

Chapter 0

Acknowledgements

It is my pleasure to acknowledge the following individuals and institutions for their support during my Ph.D. work. First, I would like to thank Adj. Prof. Xiaolin Cheng and Prof. Jeremy C. Smith for their technical guidance during my research, for enhancing my scientific understanding, and for providing a research environment that transformed me into an independent thinker. I thank the Faculty of Mathematics and Computer Science at Heidelberg University, the Heidelberg Graduate School of Mathematical and Computational Methods for the Sciences (HGS MathComp), the Interdisciplinary Center for Scientific Computing, and the Oak Ridge National Laboratory (ORNL) for their assistance throughout my work. In particular Julia Cooper, for organizing all my visits to the ORNL and Ellen Vogel for the help in Heidelberg.

This research would have not been possible without the immense computational resources provided by several supercomputing sites. For their help with the computing infrastructure, I thank Bogdan Costescu, Michael Galloway, Steve Moulton, Dr. Sabine Richling, and Steffen Hau. For the computational time, I thank the Baden-Württemberg GRiD, the National Energy Research Scientific Computing Center (NERSC) for the "Hopper" and "Franklin" resources, the National Institute for Computational Sciences (NICS) for "Kraken", and the National Center for Computational Sciences (NCCS) for the "Jaguar" and "Titan" resources. For their financial support, I thank the Baden-Württemberg Stiftung, the Oak Ridge Institute for Science and Education, and the HGS MathComp.

I thank my colleagues at the Computational Molecular Biophysics Group in Heidelberg and at the Center for Molecular Biophysics at ORNL for their support and encouragement. In particular my wonderful friends and colleagues Lipi Thukral, Mithun Biswas, Tomasz Bereźniak, Ricky Nellas, Pavan Ghatty, Farooq Kiani, and Razif Gabdoulline for helping me to grow and for being around in good and difficult times. I thank Dr. Tongye Shen for organizing the unforgettable weekend outdoor adventure trips and creating some of the adrenaline-rich experiences. He gave me a glimpse in the beautiful nature of the U.S. national parks and provided the required distraction and refreshing balance to science. I also thank my friend Richard Higginbotham for sharing the warmth of the southern hospitality. Finally, I would like to express gratitude to the unending support and encouragement of my family.

Oak Ridge, December 2012 *Emal M. Alekozai*

Search for the truth is the noblest
occupation of man; its publication is a duty.

Madame de Stael 1766-1817

Chapter 0

List of Publications and Achievements

The publications marked with (*) are presented in this thesis.

1. Pavan K. GhattyVenkataKrishna, **Emal M. Alekozai**, Greg T. Beckham, Roland Schulz, Michael F. Crowley, Edd C. Uberbacher, Xiaolin Cheng. "Initial Recognition of a Cellodextrin Chain in the Cellulose-Binding Tunnel May Affect Cellobiohydrolase Directional Specificity". *Biophysical Journal, Volume 104, Issue 4, 904-912, 19 February 2013.*

2. (*) **Emal M. Alekozai**, Pavan K. GhattyVenkataKrishna, Edward C. Ueberbacher, Michael F. Crowley, Jeremy C. Smith, Xiaolin Cheng. Simulation Analysis of the Cellulase Cel7A Carbohydrate Binding Module on the Surface of the Cellulose Iβ. *Submitted to "Cellulose".*

3. (*) **Emal M. Alekozai**, Xiaolin Cheng, Jeremy C. Smith. "Effect of the Linker Length and Stiffness on the Interaction of Cellulase Cel7A with the Cellulose Iβ Crystal Model". *To be submitted.*

While working for my dissertation, I was awarded the unique possibility to present my research output at the following selected international conferences:

- Awarded a platform session talk "Analysis of the Cellulose-Cellulase Interaction" at the Biophysical Society 54. Annual Meeting 2010 (San Francisco, USA).

- Outstanding young researcher award talk "Multilevel Enhanced Sampling of Cellulose-Cellulase Interaction" at the international workshop "From Computational Biophysics to Systems Biology 2011" (CBSB11) organized by the Juelich Supercomputing Centre, Michigan Tech, and the German Research School for Simulation Sciences. The symposium was dedicated to the 90th birthday of Prof. Harold Scheraga (Cornell University).

To my family and close friends

It would be possible to describe most things scientifically, but it would make no sense; it would be without meaning, as if you described a Beethoven symphony as a variation of wave pressure.

Albert Einstein 1879-1955

Contents

I would have written a shorter letter, but I did not have the time.

Blaise Pascal 1623-1662

The Renewable Fuel Standard requires transportation fuel to contain 36 billion gallons of biofuels by 2022. In addition, at least 16 of the 36 billion must be cellulosic biofuels. But the U.S. produces only one-third of that amount today.

U.S. Report GAO-11-513

Chapter 1

Introduction

1.1 Importance of Non-Food Cellulose Waste Derived Fuels

Biomass was the primary energy source for most civilizations from the invention of fire until the middle of the nineteenth century [1]. During the last two thousand years its consumption increased by a factor of 20 [1]. With the development of the steam engine at the end of the eighteenth century coal replaced biomass, which was then crowded out in the twentieth century by oil and gas. Today biomass only accounts for 10 % of the world's primary energy consumption. In industrialized countries 22 % and in developing countries 14 % of the total energy is consumed by road vehicles, they consume half of the world's fuel production [1]. They are responsible for 13 % of the greenhouse gas emissions and 19 % of the world's CO_2 output [1]. The current trend indicates that the amount of automobiles are doubling every 30 years [2]. It is expected that by the year 2050 the world population will reach over nine billion people. The current energy production cycle is essentially built on fossil fuels. This heavy dependence on nonrenewable fossil fuels is not a sustainable solution. For the further development of humanity, the production of food and energy has to become more efficient and has to come out of sustainable sources (Figure 1.2). Non-food biomass derived fuels, in particular from waste cellulose, can be a promising solution. Using biomass derived liquid fuels to run engines is however not a modern idea. In the nineteenth century Rudolf Diesel designed his eponymous engine to run on peanut and vegetable oil [3].

Cellulose was discovered and isolated from green plants, by Anselme Payen, over 150 years ago [4]. It is the most abundant biopolymer in earth ($\approx 10^{11}$ dry tons/year) [4–9]. It is synthesized by a great diversity of living organisms like trees, plants, bacteria, fungi, algae, and even some animals [9]. Even *cyanobacteria*, one of the most ancient forms of life on earth, synthesizes cellulose [9]. The abundance of cellulose makes it a potential renewable energy source. Cellulose consists of glucose subunits which can be unlocked and fermented to produce bioethanol. The glucose

C-H bonds of the cellulose fiber are where the useful energy is stored (Figure **??**). The aim of biofuels is to reach a fast turnover (< 1 year) of sunlight into fuel. The main reasons in favor of non-food cellulose based biofuels are [1, 3, 10–17]:

1. **Environmental importance:** Cellulose is the most abundant biopolymer on earth. CO_2 produced by burning cellulose bioethanol is accumulated by plants from the air. Compared to fossil fuels it does not generate a net unbalance of greenhouse gas and can help to downplay the growing effect of the global climatic changes.

2. **Improved air quality:** Compared to fossil fuels, biofuels have less impurities like sulfur or oxides, which are the main cause of bad air quality in large cities like Los Angeles, Beijing, Mexico City or São Paulo. For example in São Paulo 50 % of the fuel used by automobiles was replaced by biofuels, which improved the air quality remarkably [18–20].

3. **Lower impact on food production:** Biofuels using cellulose waste rather then food crops as a source of sugar can have a lower negative impact on the food production. Each year alone 40 million tonnes of inedible plant material like wood shavings from logging, corn stover (the stalks and leaves) and wheat stems are produced, most of which is thrown away [12].

4. **Exhaustion of fossil fuel resources:** It is anticipated that the worldwide fossil fuel consumption is increasing rapidly. The current fossil fuel resources are estimated to last for approximately 40 years. The production of ethanol from cellulosic biomass on commercial scales can help to reduce the dependence of fossil fuels and satisfy the increasing energy demand [20, 21].

5. **Provides energy supply security:** Current fossil fuel resources are limited to unstable and uncertain politically regions of the world. Cellulose in contrast is a domestic product in most countries and can be grown in most climate areas. It might therefore reduce the political dependence from foreign countries.

6. **Provides economic stability:** Our modern society depends highly on mobility, this strong dependence together with the unpredictable price changes of fossil fuels can influence the economical and political action scope of governments.

7. **Reusing todays infrastructure:** The existing transportation technology and infrastructure like engines and gas station are optimized during the last centuries

for liquid fuels like gasoline and diesel. Biofuels are also liquid and do not require major changes. Other renewable energy sources like fuel cells, electrical vehicle engines or gas engines would require a completely different infrastructure and technology. Liquid fuels seem to be in general more convenient for use in current vehicles then other technologies.

The political decision makers have realized this and have set following ambitious aims (Figure 1.1):

- To promote biofuels by 2011 governments in at least 17 countries have introduced targets requiring the blending of 5 to 10 % of bioethanol to gasoline, a mixture that most vehicles can run on with ease [1].

- The European Union (EU) subsidies for biofuels are nearly $5 billion a year [1]. Europe's annual requirements for transport fuels are 370 billion liters [14]. The EU aims to derive 10 % of its transportation fuels from biofuels by 2020 and 25 % by 2030 [1, 22].

- The U.S. government invested $800 million directly into its biomass program. The total subsidies for biofuels are nearly $7 billion a year [1]. The long-term goal is to use 136 billion liters (36 billion gallons) of biofuels for transport by 2020. The U.S. aim, is to supply 30 % of the 2004 motor gasoline demand with biofuels by the year 2030, which translates roughly into 230 billion liters (60 billion gallons) per year [23].

- China, the world's third largest bioethanol producer, provides around $2 billion in direct subsidies for renewable energy and has an ethanol blending target of 10 % [1].

- India, the second most populated country in the world, has set an ambitious target to meet 20 % of its fuel demand with biofuels by 2017 [1].

- "Bloomberg New Energy Finance" estimates that in 2009 governments provided at least $43 billion subsidies to the renewable energy and biofuels industries [1].

The worldwide production of ethanol has reached 22 billion liters in 2007 and should reach 47 billion liters in 2015 and should require approximately six million hectares of land [10, 20, 24]. If the scientific and political problems are solved, the International Energy Agency estimates that biofuels could meet 27 % of the global transportation fuel demand by 2050 [1].

One of the main problems to achieve these ambitious goals is that cellulose is a very stable molecule. The cellulose glucose subunits are held together by a glycosidic bond, which has a half life time of 5 to 8 million years. Cellulase enzymes, in particular Cel7A, are in many ways “protein machines” they can be promising candidates in the cellulose degradation. Using enzymes from domesticated yeast strains to convert plant derived sugars into ethanol is nothing new, mankind is using this strategy for brewing alcoholic drinks for thousands of years. Cel7A is a multi-domain enzyme consisting of a carbohydrate-binding module (CBM) and a catalytic domain (CD), joined by a linker peptide (Figure 1.3). The interaction of Cel7A with cellulose fiber is not understood in detail at the atomic level. For the research presented in this thesis parallel simulations on the some of the world most powerful supercomputers were performed to give new insights at this level. Methods from computer science, mathematics, and biophysics were utilized for the analysis of this simulations.

1.2 Thesis Outline

Three major steps are required to reach the goal of cost-effective biofuel production from non-food cellulose waste.

1. Better understanding of the interaction of the cellulose enzyme Cel7A with the cellulose fiber.

2. Designing more efficient cellulase enzymes.

3. Developing next generation energy plants which serve as improved substrates and are easier to hydrolyze with enzymes.

The aim of this thesis is to contribute in understanding the enzymatic cellulose degradation to overcome the biomass recalcitrance problem (step 1). The new insights obtained from this study, can in the long term assist in designing more efficient cellulase enzymes and better cellulose substrates (step 2 and 3).

The thesis is organized as follows. First, Chapter 2 introduces the biochemistry and -physics background, in particular the cellulose fiber and the Cel7A cellulase enzyme. To understand the Cel7A-cellulose interaction computer simulations are performed. Computer simulations additionally allow to derive, model, and investigate optimized artificial cellulase enzyme which do not yet exist in nature and which first have to be genetically engineered. The modeling and simulation framework, like the

Brownian dynamic (BD) and Molecular dynamic (MD) methodology, are presented in Chapter 3. Interesting dynamical process of the Cel7A-cellulose interaction occurs on the time scale range from nano- to milliseconds. The Cel7A-cellulose system has approximately 350.000 atoms. With present day supercomputers it is still challenging to reach convergence on the millisecond time scale for such huge systems. To tackle this problem multiscale and enhanced sampling strategies can be used (Chapter 4). The simulations protocol consists of two steps. BD simulations are performed to get insights for the encounter process on the millisecond time scale. Structures from the BD simulations are then refined using MD simulations to get a better understanding of the local interactions. The relevant Cel7A-cellulose configuration space can be sampled with a single BD simulation which is several milliseconds long, even on a supercomputer this would take too long and is therefore not practical. Instead multiple short BD simulations were performed to sample the relevant space. The statistical informations of this simulations are combined by using a Markov state model (MSM).

The Cel7A enzyme consists of the CBM domain, the CD domain, and the linker peptide. All three parts are thought to be important for the interaction of Cel7A with the cellulose fiber. The Chapter 5 asks the question:

What is the role of the CBM?

To answer this, I look at the interaction hotspots of the CBM with the fiber and determine the relevant forces at the different steps of the interaction.

Chapter 6 is dedicated to the next question addressed in this thesis.

What is the role of the linker peptide?

I derive coarse grain parameters for the linker, model the linker as a spring, and perform simulations with different spring lengths and stiffness. I analyze how the Cel7A interaction with the fiber is affected. Finally, I propose an optimal linker length and stiffness. This optimized values might assist in designing optimized cellulase enzymes.

Chapter 7 summarizes the presented results and proposes potential perspectives for future follow-up studies.

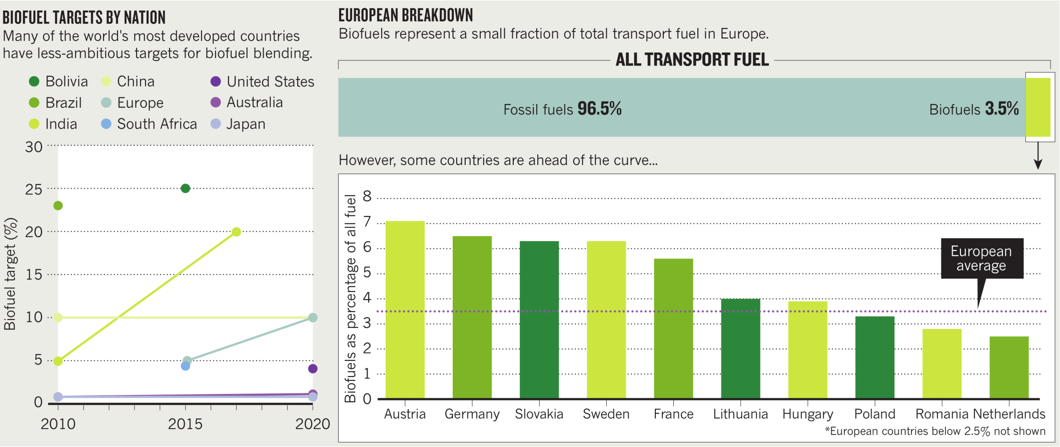


Figure 1.1: Biofuel targets by nation and overview of biofuels in transport fuel in Europe (adapted from [1]).

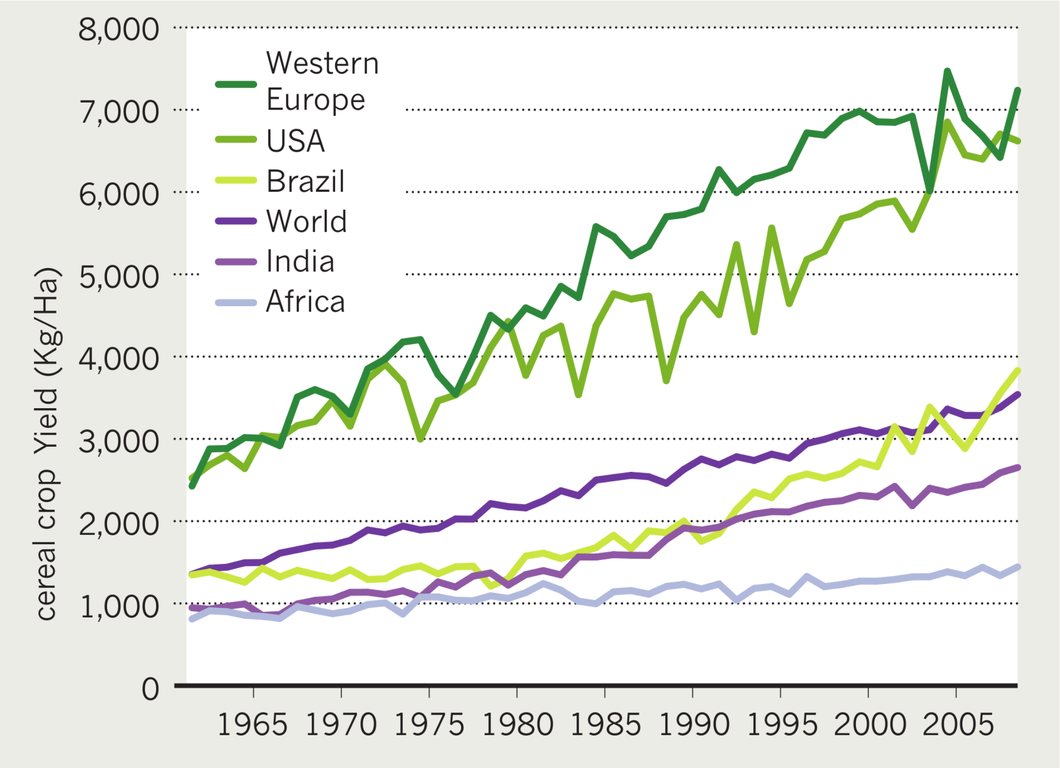


Figure 1.2: Cereal crop yields development over the last 50 years (adapted from [15]).

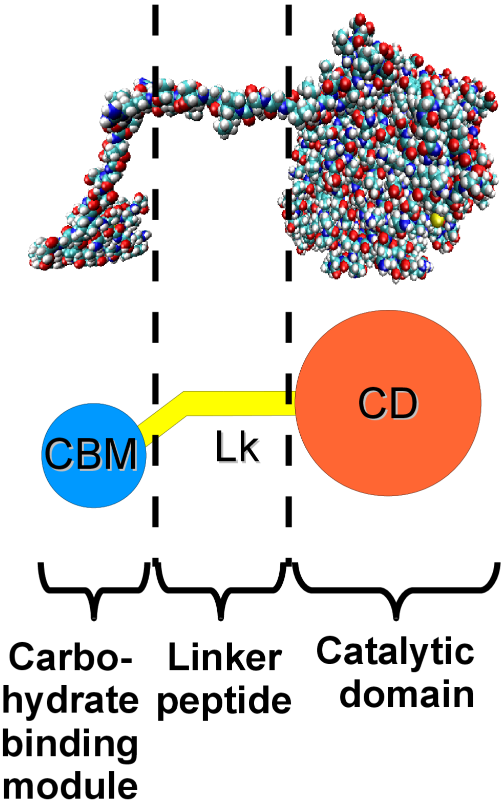


Figure 1.3: The complete Cel7A enzyme consisting of the CBM and the CD domain connected via a linker peptide.

You can plant a dream.

Anne Campbell 1992-2005

Chapter 2

Biochemistry and -physics of Cellulose and Cellulase

In this chapter the biochemical and -physical methods of the thesis on hand are described, in particular the cellulose fiber and the cellulase enzyme Cel7A.

2.1 Cellulose Fiber

2.1.1 Food Versus Fuel: 1st and 2nd Generation Biofuels

The current approach to produce biofuels, which is coined in the literature as *first generation biofuel* technologies [25, 26], is to use mainly staple food like corn, wheat, and other grains as glucose source (Figure 2.1). This approach can interfere with the food production and can have a negative impact on primary food supplies and prices [27]. Because of these concerns a better approach would be *second generation biofuels* technologies, where biofuels are produced from nonfood material and plant waste like cellulose. The main crops for first generation biofuels are corn and soya in the U.S., sugarcane in Brazil, palm-oil in south Asia, and canola in Europe. First generation biofuels do not provide a convincing solution, the arable farm land is not sufficient to cover more then 10 % of the world fuel requirements of the industrialized countries [28]. Biofuels have in general following disadvantages:

1. **Reduction forest area:** Increase use of biofuel has led to a reduction of the total world forest area.

2. **Interference with food production:** The use of available agricultural area for the biofuel production, can interfere with food production, and can lead to rising food prices.

3. **Ground and surface water polution:** Biofuel production is already at a scale that it reshapes the agriculture around the world. In the U.S. the use of corn for bioethanol has already more then tripled from 2005 to 2010. More than a third of the U.S. corn crop goes to ethanol facilities. Although corn

only consumes 10 % of the U.S. farmland (including cropland and grassland) it consumes 40 % of the fertilizer used in the U.S.. This results in an substantial increase in the amounts of fertilizer infiltrating ground and surface waters [17].

4. **Transport of biomass to biorefinery:** Bioenergy plants, such as cellulose, have a low energy density per volume and not all populated regions are suited for biofuel crops. Production areas can be scattered over a large area, the crops have to be transported over long distances to biofuel plants, which increases the production and energy costs.

The problems 1 to 3 can occur more dominant in first generation biofuels, they can be alleviated by using second generation biofuels, which use non-food cellulose waste. The fourth problem can be tackled for first and second generation biofuels by using small sized mobile production facilities near the biomass in question. Non-food cellulose can come from multiple sources like [28] (Figure 2.2):

1. **Forest waste:** Cellulose waste from saw mills, furniture, and paper industry.

2. **Forest reduction and lighting:** Forests can be lighten and trimmed regularly to reduce unproductive competition between old and new trees. Hereby cellulose can be produced without any significant influence on the forest ecosystem.

3. **Crop waste:** Crop waste like stems, blades, and leafs make 50 % of the crop. A part has to stay on the fields put the rest rotes unused.

4. **Special energy plants:** This energy plants can grow on soil which is not suitable for crop production. Examples are switchgrass, Sudan grass, Chinese silver grass, and Energy cane (a special sugarcane sort with an high cellulose proportion). Using second instead of first generation cellulose crops, like switchgrass, can even help to spare water [17].

Each year more then 100 billion dry tons of cellulose are produced. Alone in the U.S. 1.3 billion tons can be produced without reducing the crop, animal food, or other agriculture export products. This corresponds to 400 billion liter of fuel, which is half of the current gasoline and diesel consumption in the U.S. [28]. Using similar estimates the worldwide producible biomass fuel is around 5,400 to 25,000 billion barrels, which excides the world wide fuel consumption of 4,800 billion barrels [28]. The aim is to extract the energy from the cell walls, as they form up to 70 % of the plant body. The cell walls are composed of linked sugar molecules. First generation biofuels can be produced from several feedstocks, which differ in their cost and

Source	Energy ratio	Agriculture area productivity
sugarcane	8:1	2105 gallons per acre
corn	1.3:1	459 gallons per acre
switchgrass	6.4:1	500 gallons per acre

Table 2.1: Ratio of energy contained in a liter bioethanol compared to the energy used during the production process [10, 29].

efficiency. The ratio of energy contained in a liter ethanol compared to the energy used during the preparation process is shown in Table 2.1.

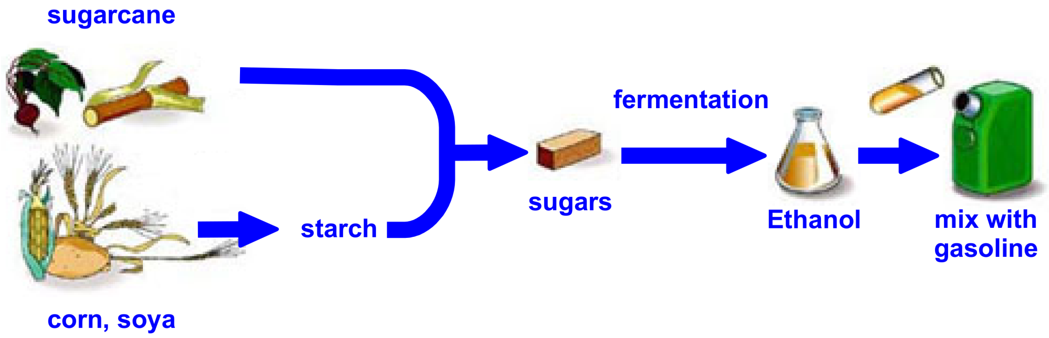


Figure 2.1: First generation biofuels use as major glucose source crop products like corn, soya bean, sugarcane, and canola. The sugars can be fermented to biofuels.

2.1.2 Molecular Structure of Cellulose

The woody material which gives plants their structure and rigidity consists of following three carbon-based polymers, which are collectively called lignocellulosic biomass: cellulose, hemicellulose, and lignin [12]. Cellulose consists of glucose bound to long chains, which are organized in crystalline microfibrils, these glucose molecules are largely insoluble. The exact chain length is unknown but single chains containing up to 14,000 glucose residues have been observed, corresponding to a fibril length of 7 μm [30, 31]. The cellulose chains are packed side by side to form microfibrils, which are $\approx$ 30 Å thick in most plants [8] (Figure 2.3). In contrast to cellulose, hemicellulose consists in addition to glucose of several other sugars like xylose, mannose, galactose, rhamnose, and arabinose. Hemicellulose consists of shorter chains and are branched, whereas cellulose are unbranched [32]. Lignin is a complex chemical compound, it is relatively hydrophobic and aromatic in nature [33]. The cellulose microfibrils are attached to hemicellulose, which is surrounded by lignin which protects the cellulose

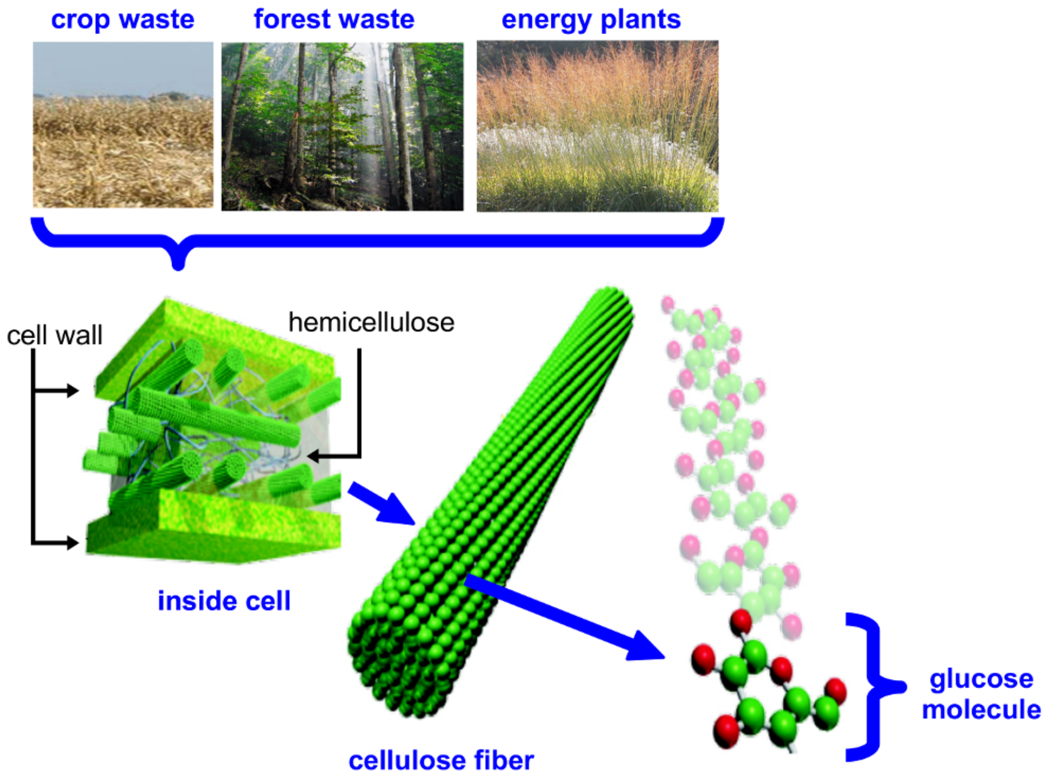


Figure 2.2: Non-food cellulose waste can come from multiple sources like crop waste, forest waste, and special designed energy plants. Cellulose consists of linked glucose sugars (adapted from [28]).

and hemicellulose [12]. The lignin builds a complex cross linked network, these strong bonds make it very difficult to penetrate the lignin [12]. The best way to break down the lignin network is to use heat and strong chemicals.

Cellulose is an unbranched chain β-D-glucose sugar units ($C_6H_{12}O_6$). Two different glucose isomers, α-glucose and β-glucose exist, in which the glycosydic hydroxyle group is located either blow or above the ring plane, respectively (Figure 2.4). The prefix "D" refers to the optical activity of the glucose, a property of rotating polarized light either to the right or to the left when the glucose is put into the light path [34]. The β-D-glucose units are linked to a cellulose chain via (1→4) glycosidic bonds. The glycosidic hydroxyl group on C_1 of one unit undergoes a reaction with the hydroxyl group on C_4 of another unit. One of the units has to turn upside down, so that the hydroxyl on C_1 is in the same plane as the hydroxyle on C_4 (Figure 2.5). During this reaction a solvent molecule (H_2O) is released. Glucose molecules have a reducing end (aldehyde group on C_1) and a non-reducing end (hydroxyl group on C_6), which gives the cellulose microfibrils a directionality [5] (Figure 2.6).

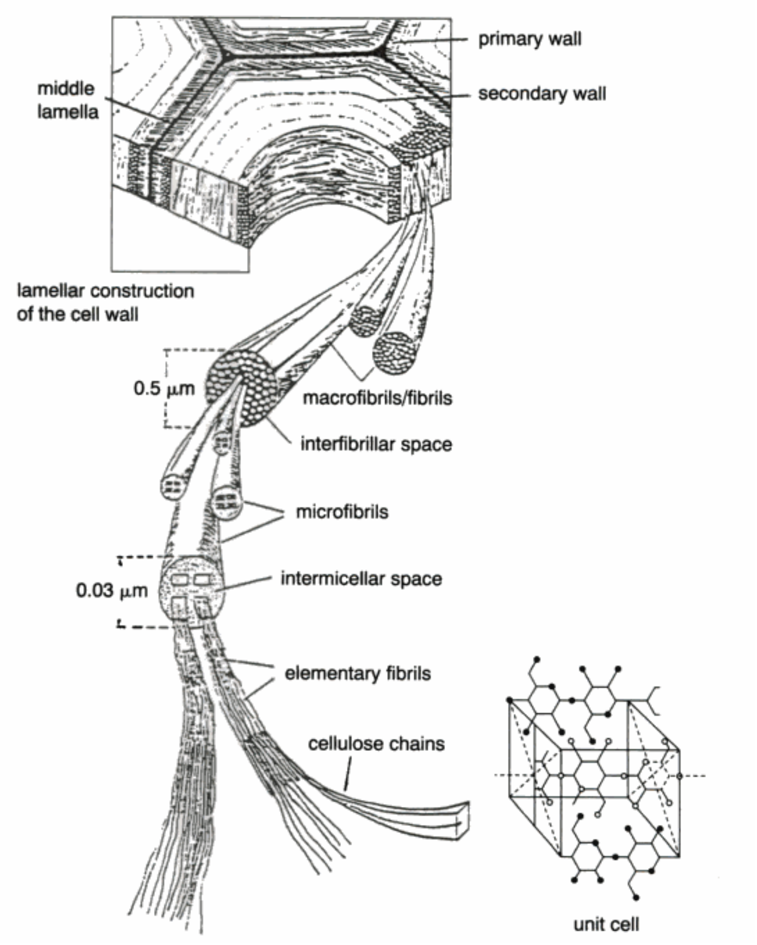


Figure 2.3: Organization of cellulose chains inside the cell wall of wood (adapted from [34]).

2.1.3 Miller Indices

A cellulose fiber consists of different fiber faces, which are better visible in the end-on view of the cellulose Iβ microfibril (Figure 2.10). To label this fiber faces the Miller index notation system is used [35]. The Miller indices are commonly used in crystallography for planes and directions in crystal lattices. Let a_1, a_2, and a_3 be the three lattice vectors pointing in the three space directions of the coordinate system. (l, m, n) defines a plane that intercepts the three vectors at points a_1/l, a_2/m, and a_3/n, or some integer multiple thereof. If one of the indices is zero, it means that the planes do not intersect that axis (the intercept is "at infinity"). Examples of

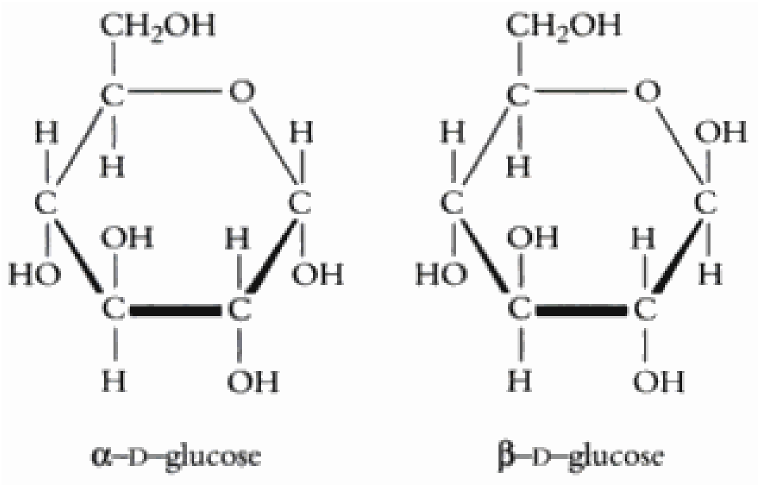


Figure 2.4: Two isomers of D-glucose. Thick lines in the plane of the ring represent the bonds facing forwards, while the thin lines represent those facing backwards. The groups attached to carbon-1 (left) are situated either below ($\alpha - \mathrm{D} - \mathrm{glucose}$) or above ($\beta - \mathrm{D} - \mathrm{glucose}$) the plane of the ring (adapted from [34]).

determining indices for the plane (1, 1, 1) and (2, 2, 1) are shown in Figure 2.7.

2.1.4 Different Types of Cellulose

The cellulose chains are closely packed too bundles of 30 to 100 chains, lying more or less parallel and form an elementary fibril. These cellulose chains are hold together by a hydrogen bond network. Six different cellulose crystalline structures I, II, III_1, III_{11}, and IV_{11} are known, depending on the location of the hydrogen bonds between and within the chains. Cellulose I, is the native cellulose form found in nature. Cellulose I can be converted to the other polymorphs [4]. A large number of intra- and intermolecular hydrogen bonds between the hydroxyl groups leads to a close packing. The smallest identical unit of the cellulose crystal is called the unit cell. Based on the packing order of the unit cell two different cellulose I types, Iα (triclinic unit cell containing one chain per unit cell) and Iβ (monoclinic unit cell containing two parallel chains per unit cell) exist [36, 37] (Figure 2.8). Native cellulose from almost every source is a mixture of Iα and Iβ, in variable proportions [8]. Cellulose produced by bacteria and algae is enriched in Iα, while cellulose of higher plants consist mainly of Iβ cellulose. Both forms Iα and Iβ are recalcitrant to hydrolysis most likely because of the enhanced inter-chain hydrogen-bonding networks [38, 39]. Microfibrils have been experimentally observed to twist, because of the small total

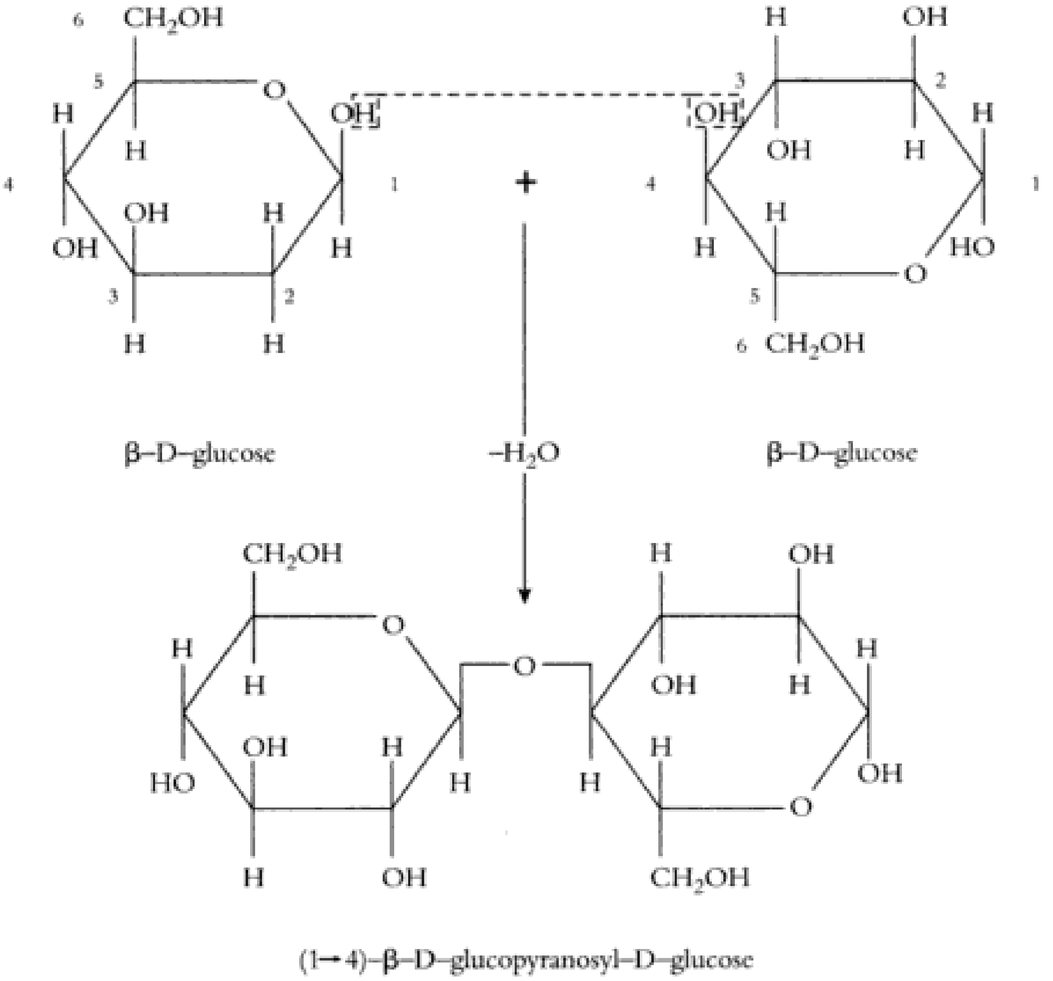


Figure 2.5: Condensation reaction between two $\beta - \mathrm{D} -$ glucose units yielding cellobiose (adapted from [34]).

length of the cellulose fiber used in this work the twist is ignored.

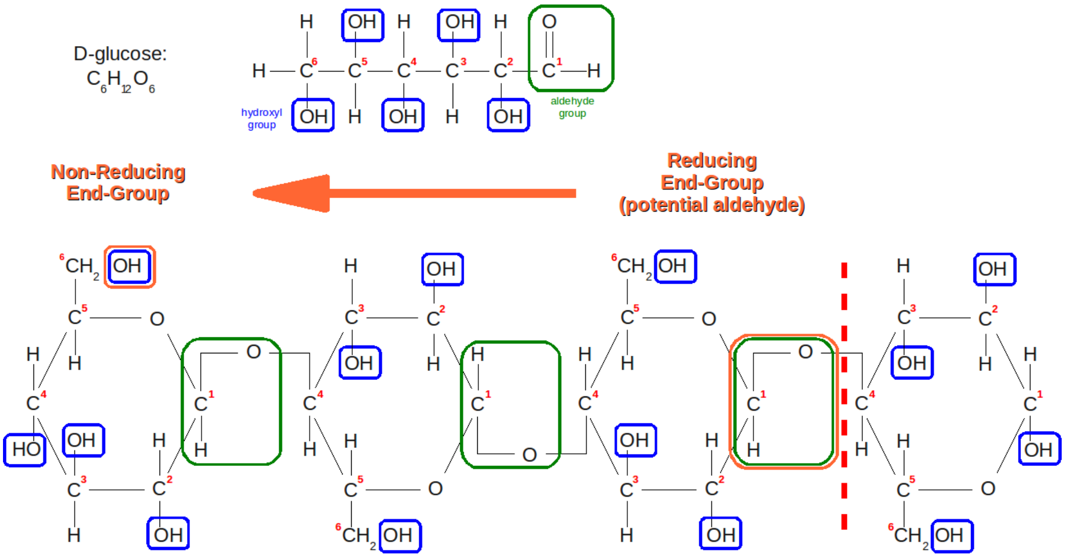


Figure 2.6: (Top left) Chemical formula of D-glucose ($C_6H_{12}O_6$). (Top right) structural formula of a single D-glucose ($C_6H_{12}O_6$) molecule. The five carbons 2, 3, 4, 5 and 6 carry an hydroxyl groups (−OH). The carbon 1 has an aldehyde functional group (−CHO). (Bottom) The chemical linkage of the oxygen (attached to C_1) and the C_4 attached to the next glucose unit can open to generate an aldehyde (−CHO). Which makes the cellulose microfibril asymmetric and allows to assign a directionality. The potential aldehyde group (at C_1) is labeled as "reducing end-group". The hydroxyl groups (−OH) on C_6 determines the "non-reducing end-group". Cel7A is a processive exocellulase which prefers to hydrolyze the cellulose chain starting from the reducing towards the non-reducing end-group.

2.1.5 Hydrophobicity of Cellulose Fiber Faces

On the different faces of the cellulose fiber different molecules are pointing out towards the cellulase enzyme, which can have an influence of the interaction of the enzyme with the cellulose fiber (Figure 2.9). The equatorial position (horizontal axis of a flat molecule) in cellulose are rich in hydroxyl groups (OH group on C_1, C_2, C_3, C_4, and C_6). The hydroxyl groups are polar and hydrogen bonding, which results in a large hydrophilic surface at the sides of the cellulose chain. The axial position (vertical axis of a flat molecule) are occupied by aliphatic hydrogens (H on C_1 to C_6). The aliphatic hydrogens are nonpolor and non-hydrogen bonding, which means that the top and bottom surface a cellulose chain is hydrophobic compared to the surface at the sides of the cellulose chain (Figure 2.9). Work using electron microscopy [40], single molecule fluorescence [41], and atomic force microscopy [42] have shown that the Cel7A CBM binds preferentially to the hydrophobic fiber faces of cellulose Iα.

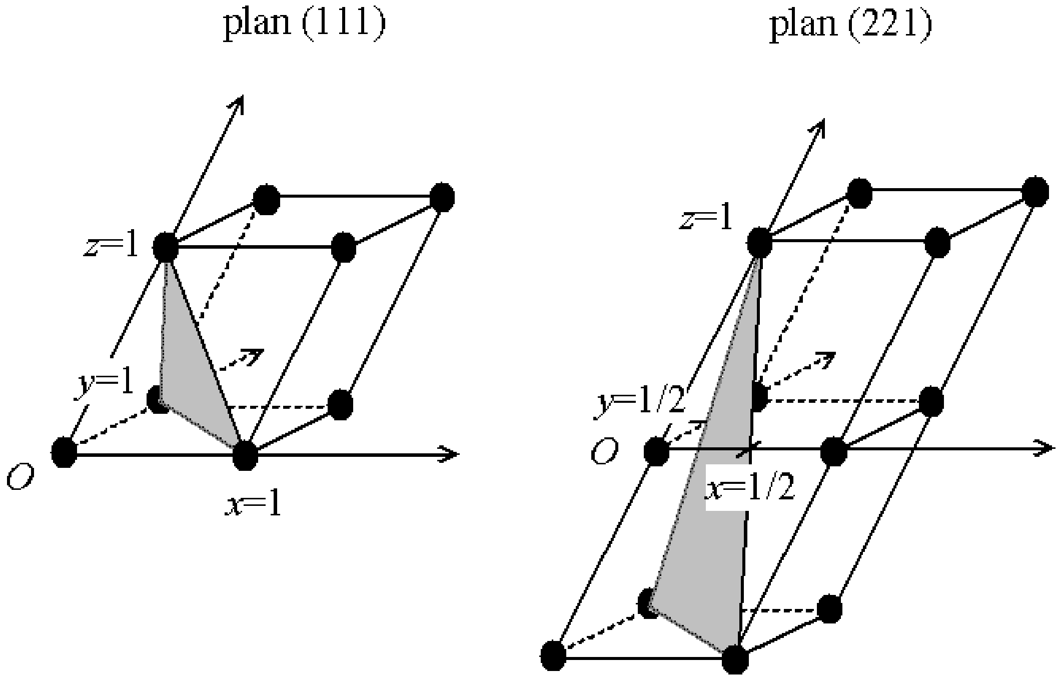


Figure 2.7: Examples of determining indices for the plane (1, 1, 1) and (2, 2, 1) (adapted from Wikipedia, Christophe Dang Ngoc Chan).

The fiber model used in this study has eight different fiber faces (Figure 2.10). The fiber faces (1, 1, 0) and (-1, -1, 0), each consist of only two chains, which are too small for a quantitative analysis, therefore they are excluded from the discussion below. Of the remaining six fiber faces, (1, 0, 0) and (-1, 0, 0) are the most hydrophobic, (0, 1, 0) and (0, -1, 0) are the most hydrophilic, (1, 1, 0) and (-1, -1, 0) being intermediate (referred to here as mixed).

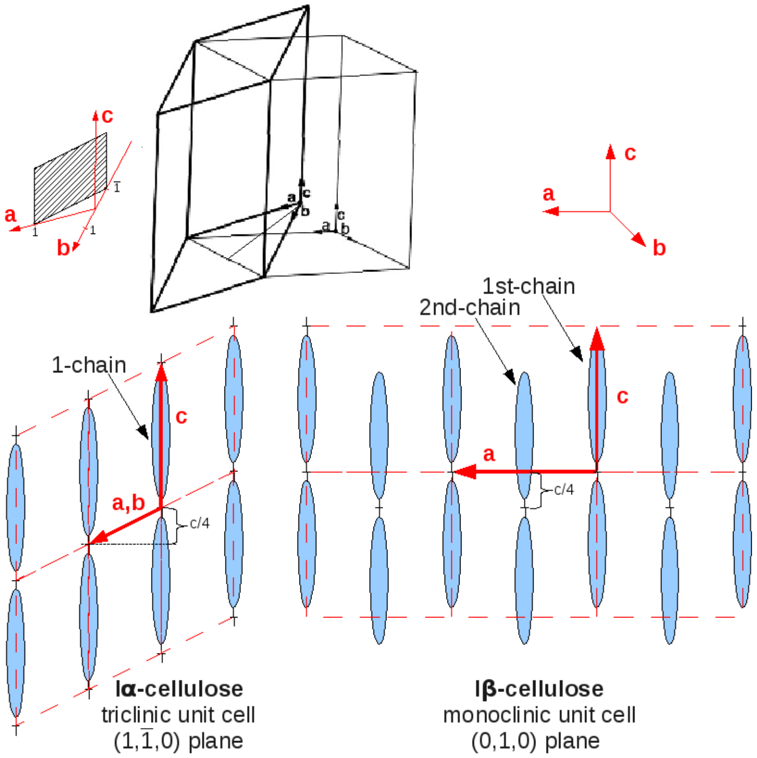


Figure 2.8: Cellulose Iα has a triclinic unit cell ($a \neq b \neq c$; $\alpha \neq \beta \neq \gamma$). Each unit cell contains one chain. Along the c axis the unit cells are shifted by $c/4$ up. Cellulose Iβ has a monoclinic unit cell ($a \neq b \neq c$; $\alpha = \gamma = 90° \neq \beta$). Each unit cell contains two parallel chains. The second chain is shifted by $c/4$.

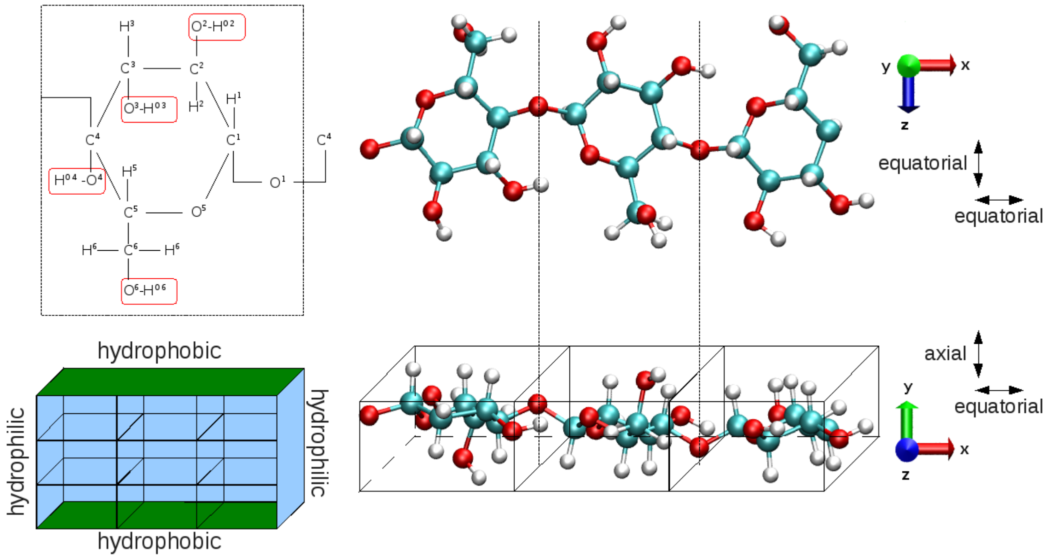


Figure 2.9: The equatorial positions in cellulose are rich in hydroxyl groups (OH group on C_1, C_2, C_3, C_4, and C_6). The hydroxyl groups are polar and hydrogen bonding, which results in a large hydrophilic surface (blue) at the sides of the cellulose chain. The axial positions are occupied by aliphatic protons (H on C_1 to C_6). The aliphatic protons are nonpolor and non-hydrogen bonding, which means that the top and bottom surface of cellulose is hydrophobic (green) compared to the surface at the sides of the cellulose chain.

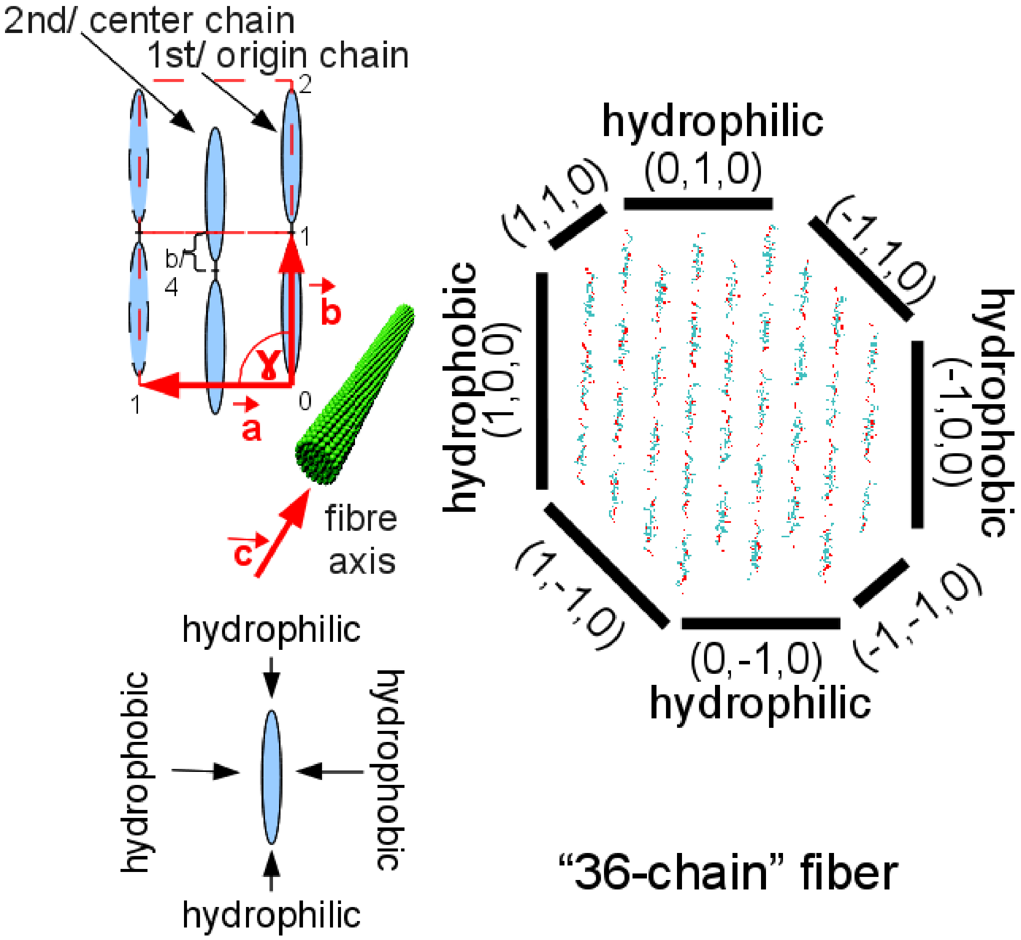


Figure 2.10: (Left top) The lattice vectors of the unit cells are shown. In cellulose Iβ each unit cell contains two parallel chains. The second chain is shifted by b/4. The view axis on the the cellulose microfibril is visualized. (Left bottom) The hydrophobicity of the different surface of the cellulose chain are shown. (Right) End-on view of the 36-chain cellulose Iβ microfibril is shown.

> I have respect for enzymes, but they can be quite fragile and they are certainly expensive.
>
> *Prof. R. Raines, Uni. Wisconsin-Madison [12]*

2.2 Cellulase Enzyme Cel7A

2.2.1 Trichoderma Reesei

Cellulose represents a significant energy reserve in the form of the chemical potential stored in its C-H and C-C bonds [**?**, 43]. This energy can be utilized by breaking the cellulose chains up into the individual sugars to produce ethanol. This is a challenging task, they are bound together via glycosidic bonds, which has a half life time of 5 to 8 million years, which makes it a very stable bond. In contrast, the half-life of amide bonds in a peptide is 125 years [44]. Plants have additionally undergone a substantial amount of evolution which helped them to develop and optimize over time complex structural and chemical defense mechanisms to prevent their deconstruction into their structural sugars by animals and microbes [23, 45]:

- The epidermal tissue of the plant body.
- The degree of lignification.
- Fermentation inhibitors which exist naturally in plant cell walls.
- The arrangement of the cellulose chains into bundles.
- The heterogeneity and complexity of cell wall constituents.
- The numerous hydrogen bonded hydroxyl groups in cellulose suggest that it is readily soluble in polar solvents. However, due to the enhanced interchain hydrogen-bonding networks [38, 39] and the high degree of crystallinity ($\approx$ 60-80 %) cellulose is insoluble in polar solvents [5, 46], which is essential for its structural role in plant cell walls.

On the other hand, in nature various animals like termite and wood-decomposing fungi happily survive on a diet of wood, indicating that evolution has also developed efficient strategies to digest wood. These organisms produce cellulose enzyme cocktails which are involved in biomass deconstruction. It is not fully known, how many enzyme types are involved and how they interact in detail [47]. Humans and animals which do not produce cellulases enzymes are therefore unable to use most of the energy contained in cellulose. In contrast, cows produce a cellulose cocktail and hence can utilize most of this energy.

Cellulases can be divided into two categories, *exocellulases* which hydrolyze cellulose chains from their termini and *endocellulases* that hydrolyze an entire glycoside linkage anywhere in the cellulose chain [48]. The cellulose fiber can have defects at non termini positions (e.g. caused by collisions or other enzymes), which then can be used by exocellulases to degrade the cellulose. The exocellulases can be further divided into two types, those that processively hydrolyze a single cellulose chain and those that remove a single chain from the chain terminus and then attack an other chain [49]. Processive cellulase enzymes, are in many ways "protein machines", they have a huge potential to efficiently hydrolyze cellulose. A promising candidate is the processive exocellulase enzyme Cel7A. It is secreted in high yields by the fungus *Trichoderma reesei* [50, 51]. *Trichoderma reesei* was discovered on the South Pacific islands during World War II, where it was eating away the garments and cotton tents of the soldiers. This fungus is also commercially exploited by manufactures of stone-washed blue jeans, laundry detergents, and paper. Cel7A is one of the most studied cellulase enzymes, it consists of two domains, a Family 1 carbohydrate-binding module (CBM) and a catalytic domain (CD), connected by a flexible linker peptide (Lk) (Figure 2.14). Cel7A consists of 497 amino-acid residues, of which the CD, the CBM, and the linker peptide consist of 434, 36, and 27 residues, respectively. A key cost factor in the conversion process from biomass to biofuel is the high cost of the cellulose enzymes [52]. 10 grams of purified *Trichoderma reesei* cost up to \$595 [1]. To make the conversion process economical more competitive and to reduce the environmental footprint higher conversion yields are required. A better understanding of the mechanism how the Cel7A interacts with cellulose is important to achieve this goal.

2.2.2 CBM

The CBM structure was derived from NMR [53]. It is wedge-shaped and has two predominant surfaces, called "bottom" and "wedge", respectively (Figure 2.13). The residues Y5, Y31, and Y32 are hydrophobic and form a hydrophobic patch on the bottom surface [54–57], while the residue Y13 is slightly buried under the wedge surface. The exact role of the CBM during the hydrolysis process of crystalline cellulose remains a debated question. The main hypotheses are:

(H1) **CBM increases local Cel7A concentration:** CBM actively binds with its hydrophobic patch to the cellulose fiber. As a consequence of this, the entire Cel7A moves closer to the fiber surface [58–60].

[1] "Worthington Biochemical Corporation" product catalog 2013, http://www.worthington-biochem.com/cel/pl.html.

(H2) **CBM is disruptive:** The CBM looses up chains from the cellulose fiber through competing with cellulose fiber intra-sheet hydrogen bonds [58, 61].

(H3) **CBM is active:** The CBM targets free chain ends and assists via its wedge face to feed the chain ends into the CD tunnel [55, 62].

2.2.3 CD

The CD is responsible for the hydrolysis reaction, it contains an active site tunnel into which a single cellulose chain can be threaded and cut into the individual sugar units [63]. The CD has a radius of $\approx 30\,\text{Å}$. The distance between tunnel entrance and exit is $\approx 50\,\text{Å}$. The residues N270 and N384 are located at the tunnel exit, and N45 is located at the tunnel entrance. The directional preference in the processive motion of Cel7A is primarily thought to arise from the structural arrangement of the CD tunnel [64]. Three glycosylation sites have been identified experimentally at the CD residues N45, N270, and N384 [50, 65]. Glycosylation patterns vary considerable depending on the growth conditions, thereby affecting enzyme activity [50, 66–71]. In the present work the glycosylation pattern of Ref. [72] was used, in which N270 and N384 have the glycosylation pattern $\text{Man}_5\text{GlcNAc}_2$ attached, and N45 is not glycosylated (Figure 2.15).

2.2.4 Linker Peptide

The linker peptide joins the CD and CBM, it has the sequence PGPSSGTT-TAPRRTTTTGPPNGGPPNG [50, 66]. The linker modeled in the fully extended form has a length of $d_{\text{LkSH}} \approx 99\,\text{Å}$. The linker consists of two regions, referred to in the literature as the hinge and stiff regions (Figure 2.14 and Figure 2.15). The linker is heavily glycosylated at the residues S4-S5, T7-T9, and T14-T17 (Figure 2.15) [50, 66]. It is based on the suggestions of Nevalainen and coworkers [73] (Figure 2.15), which is assumed to be the most common pattern in nature. The interaction of the CBM with the CD is important for the hydrolysis process, in particular the role of linker which connects both domains is not fully understood. There are several reasons why it is believed that linker is important for the Cel7A-cellulose interaction:

(H1) **Hinge:** The linker could act as a hinge between CBM and CD [74].

(H2) **Torsional leash:** The linker could act as a torsional leash between the CBM and CD [50, 74].

(H3) **Regulate CBM-CD twisting:** Several steps are required to successful hydrolyze a cellulose chain. It is speculated that the CBM actively binds to the cellulose fiber and brings via the linker the CD closer to the fiber surface. In the next step it might targets a free chain end and assists via its wedge face to feed the chain end into the CD tunnel. The relative twisting between the CBM and CD might be important herefore. The linker could regulate and stabilize the relative orientation between both domains.

(H4) **Regulate CBM-CD distance:** The linker might help to maintain and regulate the spatial distance between the CD and the CBM.

(H5) **Prevent proteolysis:** The extensive linker glycosylation might protect Cel7A from proteolysis [75, 76].

(H6) **Interaction with other enzymes:** The linker might be vital for the secretion of other enzymes within *T. reesei* [74, 77].

(H7) **Effects Cel7A activity:** Mutational experiments confirmed that the shortening or the removal of the linker peptide results in the reduction or full loss of activity in Cel7A [77].

(H8) **Spring:** The linker has the capacity to store energy in a manner similar to a compressed or stretched spring [78].

(H9) **Caterpillar like motion:** Based on experimental studies of Cel45 from *Humicola insolens* [79], it has been postulated that the linker works in a spring-like motion to enable the enzyme to move in a caterpillar like fashion along the the cellulose surface.

In summary, further work is needed to understand in detail which role the linker plays in the Cel7A-cellulose interaction.

Figure 2.11: Typical appearance of *Trichoderma reesei* fungus growing on wood (adapted from “Trichoderma, Sex, and Fuel“ Robert L. Anderson, http://mycorant.com/trichoderma-sex-and-fuel/).

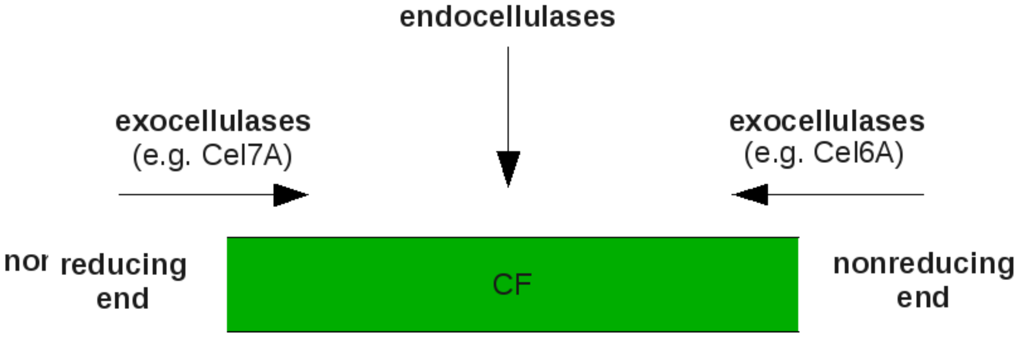


Figure 2.12: Cellulase enzymes can be divided into two categories, exocellulases and endocellulases. The exocellulases can be further divided into two types, those that possessively hydrolyze a single cellulose chain and those that remove a single chain from the chain terminus and then attack an other chain. Cel7A is a processive exocellulase which prefers to hydrolyze cellulose chains starting from the reducing end, whereas Cel6A is a processive exocellulase which prefers to start from the nonreducing fiber end.

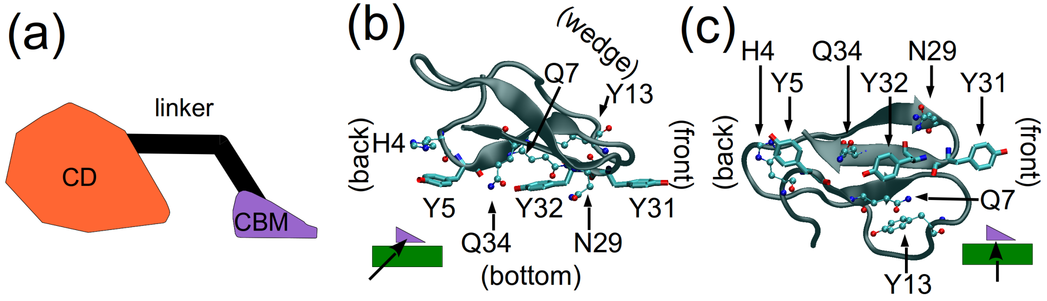


Figure 2.13: (a) The complete Cel7A. (b) Side view and (c) bottom view of the CBM with H4, Y5, Q7, Y13, N29, Y31, Y32, and Q34 highlighted

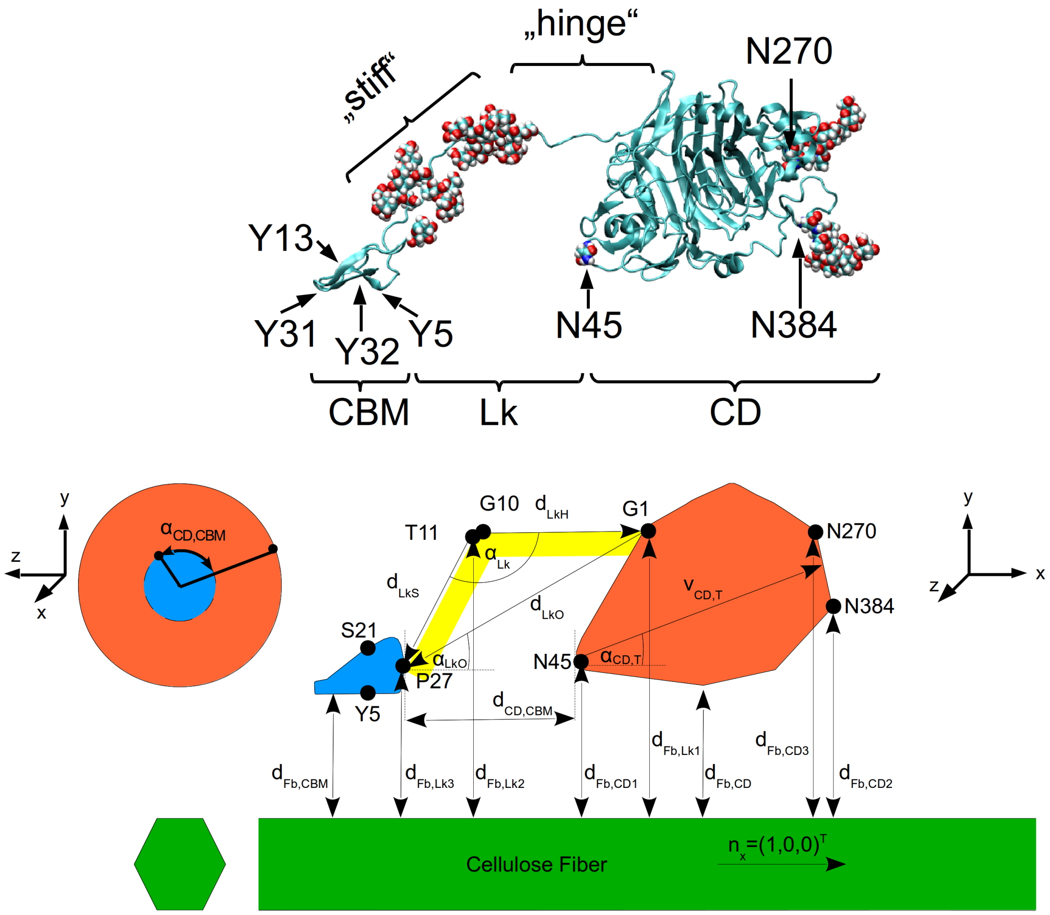


Figure 2.14: (Top) Side view of glycosylated Cel7A, the residues Y5, Y31, Y32, N45, N270, and N384 are highlighted. (Bottom) Schematic view of Cel7A, the distances and length are visualized. CBM, Lk, and CD are shown in blue, yellow, and orange colors respectively.

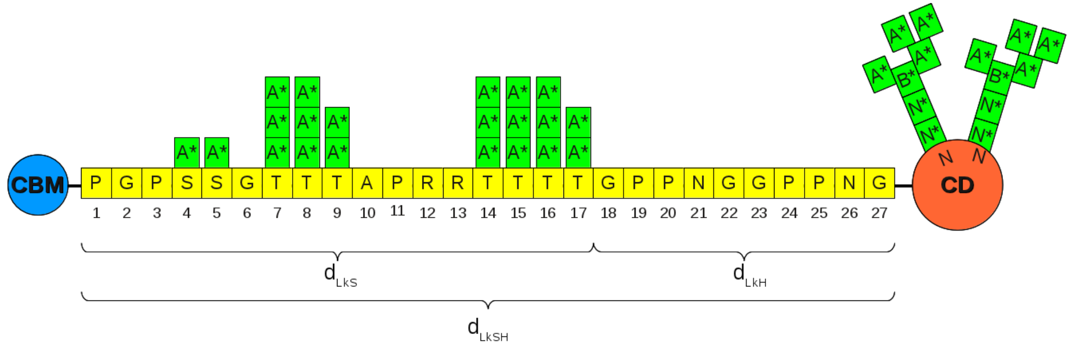


Figure 2.15: Schematic illustration of the Cel7A enzyme. The α-D-mannose is abbreviated with A*, the β-D-mannose with B*, and the β-N-acetylglucosamine with N*. Cel7A linker domain sequence with the O-glycosylation shown in yellow. The S4-S5, T7-T9, and T14-T17 residues in the linker have a O-glycosylation with a range of α-D-mannoses between 1 and 3 on each site. The CD is glycosylated at N270 and N384. The stiff region d_{LkS} is from the C_α on residue P1 to the C_α on residue T17 and the hinge region d_{LkH} is from the C_α on residue G18 to residue G27. The linker modeled in the fully extended form has a length of d_{LkSH}.

Different types of biofuels can be derived form biomass. The best replacement for petroleum is petroleum.

Prof. J. Regalbuto, University of Illinois [13]

2.3 Biofuels

The main components of plants are cellulose, hemicellulose, and lignin. Both, lignin and hemicellulose make it difficult to access the cellulose. Using high pressure and temperatures nature converted inside the earth's interior cellulose from zooplankton and algae into mineral oil fields. A similar approach can also used in refineries. Currently high cellulase enzyme loadings are required for the direct transformation of cellulose into the single glucose molecules. Current biomass conversion schemes consist of several pretreatment steps, in which a combination of mechanically grounding, heat, chemicals, and enzymatic treatments are applied. The main aim of the pretreatment steps is to enhance the enzyme activity. In a first step, the woody biomass is mechanically grounded up. In a second pretreatment step, the rigidity of the biomass is decreased by using heat and acids which rip the lignin apart and exposes the hemicellulose and cellulose. In a next step, the hemicelluloses are converted to polymers of one to ten sugars (monsaccharides and oligoscaccharides). The systematic removal of hemicelluloses during the pretreatment steps exposes the crystalline cellulose core and reduces the required cellulase enzyme loadings. Once the pretreatment steps have compromised the hemicellulose barrier an enzyme cocktail can be used to hydrolyze the crystalline cellulose cores [47]. In nature at least three different categories of enzymes are necessary to hydrolyze native cell-wall materials, hemicelluloses, and lignin [23, 80]. Another weapon in the pretreatment arsenal are ionic liquids, they are salts which become liquid at room temperature or just above [12]. They can penetrate lignin and liquefy biomass. The ionic liquid left in the sugar mixture can hinder the enzymes from functioning. The critical step is recovering the liberated biomass sugars from the ionic liquids.

The price is one of the main factors which will decide over the success of biofuels. The main competitor for biofuels are fossil fuels, which profit from over one century of research. The main investments of the refineries have been already amortized. The enzyme cocktails so far discovered are not very efficient for large scale industrial conversions. This process consumes at the moment more energy then which can be released from the sugar molecules [12]. Key for higher yields, is to understand better how the enzymatic process works. In the next step this can lead to improved enzyme cocktails which lower the conversion costs.

It doesn't matter how beautiful your theory is, it doesn't matter how smart you are. If it doesn't agree with experiment, it's wrong.

Richard Feynman 1918-1988

2.4 Neutron Scattering

2.4.1 Why Neutron Scattering?

In the context of atomic simulations neutron scattering experiments [81–89] have two main applications:

1. **Structural properties of biomolecules:** To perform atomic detailed simulations, the relative positions of the atoms of a biomolecule are required to build a three dimensional structure. The neutron absorption and scattering pattern gives insights about the atom distribution inside the biological sample.
2. **Validation of simulation results:** Most computer simulations require modeling approximations to address with the given computational resources current scientific questions (Chapter 3). One way to validate that this approximations do not significantly effect the simulation results is to compare them with neutron scattering experiments.

2.4.2 Neutron Scattering 101

Neutrons have wavelengths in the Ångström (Å) and energies in the meV range, which is of the same order as the inter atomic distance and energy of biomolecules. This makes neutrons sensitive to the amplitudes and frequencies of the atom motions in biomolecules [87]. Neutrons are characterized by their energy

$$E = \hbar\omega \tag{2.1}$$

and their momentum

$$\mathbf{p} = m\mathbf{v} = \hbar\mathbf{k}\,, \tag{2.2}$$

with the wave vector being

$$\mathbf{k} = \frac{2\pi}{\lambda}\mathbf{e}_k\,. \tag{2.3}$$

If the electromagnetic interactions are neglected the neutron has three possibilities when it passes near a sample atom:

1. Neutron *passes through* the sample, without any significant changes of its physical properties.
2. Neutron is *absorbed* in a nuclear process and its energy is absorbed by a sample nucleus.

3. Neutron is *scattered* by a sample nucleus, a energy and momentum exchange takes place:

$$\begin{aligned} \Delta E &\equiv E' - E = \hbar(\omega' - \omega) && (2.4) \\ \Delta \mathbf{p} &\equiv \mathbf{p}' - \mathbf{p} = \hbar(\mathbf{k}' - \mathbf{k}) = \hbar\mathbf{q}\,, && (2.5) \end{aligned}$$

   with the momentum transfer $\mathbf{q} \equiv \mathbf{k}' - \mathbf{k}$. Primed and unprimed quantities are after and before the scattering event, respectively. Based on the interaction of the neutron and the sample nucleus two specific forms of scattering can be distinguished. In *inelastic scattering* an exchange of momentum and energy takes place ($\Delta E \neq 0$) and during *elastic scattering* ($\Delta E = 0$) energy is conserved and only momentum is exchanged.

In experiments, two types of scattering can be observed, *coherent* and *incoherent* scattering [85]. In coherent scattering an incident neutron wave interacts with all sample nuclei in a coordinated fashion. The waves scattered from all the sample nuclei will have a definite relative phase and can thus interfere with each other. During incoherent scattering the scattered waves from different sample nuclei will have random relative phases and thus cannot interfere with each other [89].

The experimental method of choice depends on the length scale of the sample to be investigated. *Small-angle neutron scattering* (SANS) uses elastic neutron scattering at small scattering angles to investigate the structure of biological macromolecules at a scale of about 1 to 100 nm. It is often combined with *small-angle X-ray scattering* (SAXS), which uses X-rays instead of neutrons. Main advantages of SANS over SAXS are its sensitivity to light elements and the possibility of isotope labeling. Biological samples are usually dissolved in water, and in scattering experiments it is difficult to distinguish the water form the sample hydrogens. Hydrogen atoms make up about half the total number of atoms in a biomolecule and are distributed evenly throughout the biomolecule. The special behavior of hydrogen compared to deuterium in SANS is helpful in this case. The exchange of sample hydrogens with deuterium has minimal biological effect but has a significant effect on the scattering and allows to distinguish the biological sample atoms from the surrounding solvent.

2.4.3 Intermediate and Dynamic Structure Factor

From a MD simulation trajectory the position of all the atoms at each time step are known. A experimental accessible quantity in neutron scattering is the incoherent dynamic structure factor $S_{inc}(\mathbf{q}, t)$. The result of a MD simulation can be verified

by calculating the $S_{inc}(\mathbf{q},t)$ from a MD trajectory and comparing it with $S_{inc}(\mathbf{q},t)$ from a neutron scattering experiment. The calculation of $S_{inc}(\mathbf{q},t)$ from the MD trajectory is explained as described in the text below.
The dynamics of a system of n isotropically pure scattering atoms can be described by a space and time dependent correlation function $\tilde{g}(\mathbf{r},t)$ which was introduced by van Hove [88–90]:

$$\tilde{g}(\mathbf{r},t) \equiv \frac{1}{n} < \sum_{k=1}^{n}\sum_{l=1}^{n} \delta(\mathbf{r} - \Delta\mathbf{R_{kl}}(t,t_0)) >_{t_0} \tag{2.6}$$

$$\text{position of atom } k \text{ at time } t \quad : \quad r_k(t) \tag{2.7}$$

$$\text{position of atom } l \text{ at time } t_0 \quad : \quad r_l(t_0) \tag{2.8}$$

$$\text{distance between atom } l \text{ and } k \quad : \quad \Delta\mathbf{R_{kl}}(t,t_0) \equiv \mathbf{r}_k(t) - \mathbf{r}_l(t_0) \tag{2.9}$$

$$\text{Kronecker delta [91,92]} \quad : \quad \delta(a-a_0) \equiv \begin{cases} 1 & \text{if } (a-a_0)=0 \\ 0 & \text{else} \end{cases} \tag{2.10}$$

The brackets $< .. >$ donate an average over all possible starting times t_0 for observing the system. This is equivalent to an average over all the possible thermodynamic states of the sample [85]. $\tilde{g}(\mathbf{r},t)$ is proportional to the probability of finding an atom k at a position $\mathbf{r}$ at time t, given that there was a particle l at the origin $\mathbf{r}$ at time t_0. The correlation function $\tilde{g}(\mathbf{r},t)$ can be split into two terms, usually called the "self" part $\tilde{g}_s(\mathbf{r},t)$ and the "distinct" part $\tilde{g}_d(\mathbf{r},t)$. The self-correlation part has the following shape ([82], p. 233):

$$\tilde{g}_s(\mathbf{r},t) = \frac{1}{n}\sum_{k=1}^{n} < \delta(\mathbf{r} - \Delta\mathbf{R_{kk}}(t,t_0)) >_{t_0} \tag{2.11}$$

The self-correlation part $\tilde{g}_s(\mathbf{r},t)$ can be rescaled by the scattering length b_k of atom k. The rescaled self-correlation function $g(\mathbf{r},t)$ has the following shape:

$$g(\mathbf{r},t) \equiv \frac{1}{n}\sum_{k=1}^{n} b_k^2 < \delta(\mathbf{r} - \Delta\mathbf{R_{kk}}(t,t_0)) >_{t_0} \tag{2.12}$$

The Fourier transformation from $\mathbf{r} \rightarrow \mathbf{q}$ of $g(\mathbf{r},t)$ gives the so-called intermediate scattering function $I(\mathbf{q}, t)$:

$$\begin{aligned}
I(\mathbf{q}, t) &\equiv FT_{\mathbf{r},t\rightarrow\mathbf{q},t}(g(\mathbf{r}, t)) \tag{2.13}\\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathbf{dr} e^{-i\mathbf{qr}} g(\mathbf{r}, t) \tag{2.14}\\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathbf{dr} e^{-i\mathbf{qr}} \left(\frac{1}{n}\sum_{k=1}^{n} b_k^2 \left\langle \delta\left(\mathbf{r} - \Delta\mathbf{R_{kk}}(t, t_0)\right)\right\rangle\right) \tag{2.15}\\
&= \frac{1}{n}\sum_{k=1}^{n} b_k^2 \left\langle \underbrace{\frac{1}{2\pi}\int_{-\infty}^{\infty} \mathbf{dr} e^{-i\mathbf{qr}} \delta\left(\mathbf{r} - \Delta\mathbf{R_{kk}}(t, t_0)\right)}_{=FT_{\mathbf{r},t\rightarrow\mathbf{q},t}(\delta(\ldots))} \right\rangle \tag{2.16}\\
&= \frac{1}{n}\sum_{k=1}^{n} b_k^2 \left\langle e^{-i\mathbf{q}\Delta\mathbf{R_{kk}}(t,t_0)} \right\rangle \tag{2.17}\\
&= \frac{1}{n}\sum_{k=1}^{n} b_k^2 \left\langle e^{-i\mathbf{q}\mathbf{r}_k(t)} e^{i\mathbf{q}\mathbf{r}_k(t_0)} \right\rangle \tag{2.18}
\end{aligned}$$

The following definition of the Fourier transformation [91] of the function $f(a)$ from space a to space b was used:

$$\begin{aligned}
\text{definition} \quad &: \quad F(b) = FT_{a\rightarrow b}\left(f(a)\right) \equiv \frac{1}{2\pi}\int_{-\infty}^{\infty} da e^{-iba} f(a) \tag{2.19}\\
\text{example [93]} \quad &: \quad f(a) = \delta(a - a_0) \Rightarrow FT_{a\rightarrow b}\left(\delta(a - a_0)\right) = e^{-iba_0} \tag{2.20}
\end{aligned}$$

The Fourier transformation from $t \rightarrow \omega$ of $I(\mathbf{q}, t)$ gives the so-called dynamic structure factor $S(\mathbf{q}, t)$:

$$\begin{aligned}
S(\mathbf{q}, t) &\equiv FT_{\mathbf{q},t\rightarrow\mathbf{q},\omega}(I(\mathbf{q}, t)) \tag{2.21}\\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathbf{dt} e^{-i\omega t} I(\mathbf{q}, t) \tag{2.22}\\
& \tag{2.23}
\end{aligned}$$

The dynamic structure factor is just a double Fourier transform of a rescaled self correlation function $g(\mathbf{r}, t)$:

$$S(\mathbf{q}, t) \equiv FT_{\mathbf{q},t\rightarrow\mathbf{q},\omega}(FT_{\mathbf{r},t\rightarrow\mathbf{q},t}(g(\mathbf{r}, t))) \tag{2.24}$$

The Fourier transformation is a global and information conserving transformation [94, 95]. Therefore $S(\mathbf{q}, t)$ and $g(\mathbf{r}, t)$ are equivalent descriptions of protein motion. Due to the high incoherent scattering length of the hydrogen atoms in neutron scattering experiments one can neglect the contribution of coherent scattering. The total structure factor can be simplified to $S(\mathbf{q}, t) \approx S_{inc}(\mathbf{q}, t)$. In this case, in the above formulas the total scattering length b_i has to be replaced by the incoherent scattering length $b_{i,inc}$.

2.5 Dipole-Dipole Interaction

Dipole-dipole interaction are a special type of interaction between molecules. An illustrative example how vital the dipole-dipole interaction is to the human health, is the formation of red blood cells. They consist of to alpha chains, two beta chains, and a heme group. For the folding of the alpha and beta chains a series of dipole-dipole interactions is required. Any mutation that destroys the dipole-dipole interaction prevents them from forming properly and impairs their ability to transport oxygen to the tissues, which can even lead to death.

Biomolecules contain positive and negative charges. Due to the permanent non-uniform distribution of the charges on the various atoms some parts of the biomolecule can be more positive and others more negatively charged. This charge separation can be mathematically described by an electric dipole moment μ, which is a vector quantity pointing from the negative charge $\delta-$ towards the positive charge $\delta+$. The magnitude is equal to each charge times the separation between the charges. The unit of μ is the debye (symbol D). Biomolecules with an dipole are polar. This charge separation inside the biomolecule generates an electric field around the biomolecule. Different biomolecules with a permanent dipole (e.g. CBM domain and the cellulose fiber) can interact with each other via their electric field. This type of interaction is called dipole-dipole interaction. During the dipole-dipole interaction the two dipoles orient such that the total energy of the complete system is minimized. The dipole-dipole interaction of the CBM with the cellulose fiber would favor the CBM orientation relative to the fiber, in which the front side of the CBM would be facing towards the reducing fiber end (Figure 2.16).

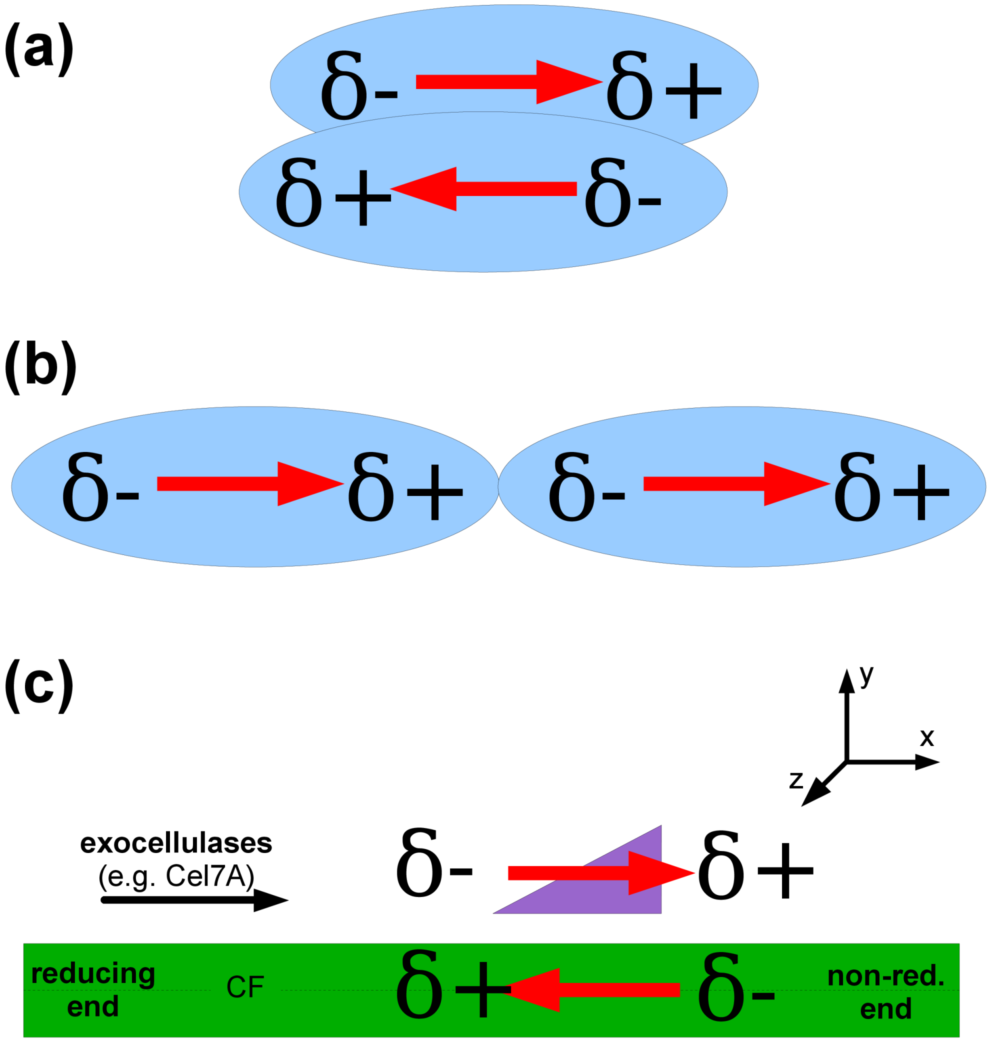


Figure 2.16: Biomolecules with a permanent dipole (shown as red arrows) attract each other, due to their partial charges. The two relative biomolecules orientations (a) and (b) would lead to a decrease of the total energy. The dipole-dipole interaction of the CBM with the cellulose fiber would favor the CBM orientation relative to the fiber, in which the front side of the CBM would be facing towards the reducing fiber end (c).

There's no sense in being precise when you don't even know what you're talking about.

John von Neumann 1903-1957

Chapter 3

Modeling of Cellulase-Cellulose Interaction

Full-atomic computer simulations allow to bridge the gap between experiments and theory for the Ce7A-cellulose system. Here two different computer simulation methods Brownian (BD) and molecular dynamic (MD) [96–101] simulations are presented.

3.1 BD Simulations

The cellulose fiber, the CBM, the CD, and the Cel7A were each modeled as an atomic-detailed rigid body Brownian particle, with each particle able to translate and rotate during the simulation. The force field used in the BD simulation consists of a combination of electrostatic (ES) and van der Waals (VDW) interaction terms. For testing purposes additional electrostatic desolvation (ED) and hydrophobic desolvation (HD) terms were also included. A detailed description of these two terms is given in Ref. [102–105]. Neither the electrostatic nor hydrophobic desolvation term significantly influences the CBM-fiber encounter process. Hence, the ES and VDW terms are sufficient to model the relevant features of the CBM-fiber encounter process while at a much lower computational cost. The BD interaction model presented here provides a balance between accuracy and speed.

To speed up ES interactions between the BD particles with high accuracy, the effective charge method (ECM) [105, 106], was used to derive charges that represent the external ES potential in a uniform dielectric medium. In the first step, the solution of the Poisson-Boltzmann equation (PBE) [105, 107–110], was obtained for each BD particle, and test charges were then assigned to each BD particle. Based on the test charges, effective charges were fitted to reproduce the ES potential of the molecule computed by solving the PBE. The linearized PBE was solved using the Adaptive Poisson-Boltzmann Solver program (APBS) [111], with a single-point multigrid method without focusing [111], with a mesh domain length of 270 Å x 270 Å x 270 Å and a grid spacing of 0.7 Å for the cellulose fiber and with a mesh domain length of 100 Å x 100 Å x 100 Å and a grid spacing of 0.5 Å for the CBM. The ionic strength was set to 150 mM and the temperature was set to 298.15 K. The relative dielectric constant for the solvent was set to 78.54 [112, 113], the dielectric constant for the CBM set to 4 and that for the cellulose fiber set to 6 [114, 115].

For solving the PBE the partial charges and atomic radii were assigned to the CBM and fiber using the program PDB2PQR [116]. The partial charges for the CBM and the cellulose atoms were taken from the CHARMM23 parameter set [117], and the CSFF force field for carbohydrates [118, 119], respectively, leading to a total charge of zero for the CBM and the 36-chain fiber. The boundary conditions were set to “Multiple Debye-Hückel” as reported in Ref. [111]. For the CBM, on each amino acid test charges of -0.5 e and 0.5 e were placed on the backbone oxygen and nitrogen atoms, respectively.

The cellulose fiber consists of repeating glucose units and is therefore highly symmetric. For the fiber models 2880 test charges were placed on the ring oxygen atom (-0.5 e) and the third ring carbon atom (0.5 e) of the surface chains. The effective charges were fitted to reproduce the ES potentials over distances 3-30 Å from the van der Waals surface of the BD particle.

To model the VDW interactions between the BD particles a Lennard-Jones like potential was used. For computational efficiency values were mapped for each BD particle on a cubic grid

$$E_{vdw}(\mathbf{k}) = \sum_{i=1}^{n} 4\epsilon_{ij} \left\{ \frac{\sigma_{ij}^{12}}{a_i^{12} + |\mathbf{k} - \mathbf{l}_i|^{12}} - \frac{\sigma_{ij}^{6}}{a_i^{6} + |\mathbf{k} - \mathbf{l}_i|^{6}} \right\}, \tag{3.1}$$

where k is a point on the grid, l_i the position of atom i, $|k - l_i|$ the distance between the center of the two atoms i and j (represented by the closest grid point k) in different BD particles, ϵ_{ij} the depth of the potential well, σ_{ij} the distance at which the inter-particle potential is zero. a_i is the VDW radius of atom i, and the terms a_i^6 and a_i^{12} remove the singularity at $|k - l_i| = 0$. To further increase the computational efficiency the Lorentz-Bertelot rules [120, 121] were used:

$$\epsilon_{ij} = \sqrt{\epsilon_{ii} \cdot \epsilon_{jj}} \approx \epsilon_{ii}, \tag{3.2}$$

$$\sigma_{ij} = \sqrt{\sigma_{ii} \cdot \sigma_{jj}} \approx \sigma_{ii}. \tag{3.3}$$

The values of σ_{ii} and ϵ_{ii} were obtained from the parameter set of the OPLS force field . A mesh domain length of 220 Å x 110 Å x 110 Å with a grid spacing of 0.7 Å was used for the cellulose fiber and a mesh domain length of 96 Å x 96 Å x 96 Å with a grid spacing of 0.7 Å was used for the CBM.

The trajectories of the BD particles were propagated using the Ermak-McCammon algorithm [122] with a time step of 1 ps. The free diffusion translational and rotational coefficients D_{trans}^{free} and D_{rot}^{free} for the BD simulations of the cellulose fiber, the intact Cel7A, the CD, and the CBM were determined from the MD sets (Table 5.7).

3.2 MD Simulations

The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble.

Paul Dirac 1902-1984

3.2.1 Born-Oppenheimer Approximation

The dynamics of an atomic system can be described by the time dependent Schrödinger equation

$$i\frac{h}{2\pi}\frac{\partial\psi}{\partial t} = \hat{H}\psi\,, \tag{3.4}$$

where ψ is the wave function, i is the imaginary unit, h is the Planck constant, and $\hat{H}$ is the Hamiltonian of the system (sum of potential and kinetic energy). ψ is a function of the coordinates and momenta of all the nuclei and electrons in the system, which makes the calculation of Equation 3.4 quite challenging. The calculation can be alleviated by the Born-Oppenheimer approximation which is an assumption that the electronic motion and the nuclear motion in molecules can be separated and that the total wave function ψ can be split into a nuclear ψ_n and an electronic component ψ_e:

$$\psi = \psi_n \cdot \psi_e\,. \tag{3.5}$$

It was proposed in 1927 by Born and Oppenheimer [123]. The approximation is motivated by the observation that the electron motion is much faster then the nuclei motion, therefore it is valid to assume that the electrons adjust instantly to any motion of the nuclei [124]. Using this approximation, Equation 3.4 can be separated into two less complicated parts.

3.2.2 Force Field

The large number of electrons in a biomolecule make the solution of the Schrödinger equation computationally challenging. The force field description, which includes the nuclear coordinated and incorporates the effect of the electrons via a potential function, is a computationally less expensive approach. In the thesis the CHARMM27 force field [125] was used, the potential energy is separated into bonded and non-bonded terms [88, 126]:

$$E = \underbrace{E_{\text{bonds}} + E_{\text{angles}} + E_{\text{dihedrals}} + E_{\text{improper dihedrals}}}_{\text{bonded terms}} + \underbrace{E_{\text{el}} + E_{\text{vdW}}}_{\text{non-bonded terms}} \tag{3.6}$$

$$= \sum_{\text{bonds}} \alpha_b (b - b_{eq})^2 + \sum_{\text{angles}} \alpha_\theta (\theta - \theta_{eq})^2 + \sum_{\text{dihedrals}} \alpha_\phi [1 + cos(n\phi - \phi_{eq})]$$

$$+ \sum_{\text{improper dihedral}} \alpha_\omega (\omega - \omega_{eq})^2 + \underbrace{\sum_{i<j} \frac{q_i q_j}{\epsilon r_{ij}}}_{\text{electrostatic}} + \underbrace{\sum_{i<j} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right]}_{\text{van der Waals}} \tag{3.7}$$

The individual bonded terms describe stretching of the bond length b, bending of the bond angle θ, torsion of the proper dihedral angle ϕ, with multiplicity n, and bending of the improper dihedral angle ω, as depicted in Figure 3.1. α_x and x_{eq} denote the force constants and equilibrium values, respectively. The last two sums in 3.7 describe the electrostatic and the van der Waals interaction, respectively. q_i and q_j are the charges of atoms i and j, r_{ij} is the distance between atoms i and j, and ϵ is the dielectric susceptibility. The Lenard-Jones 12-6 function is chosen to describe the van der Waals interaction, the parameters being the depth of the potential ϵ_{ij} and the collision parameter σ_{ij}.

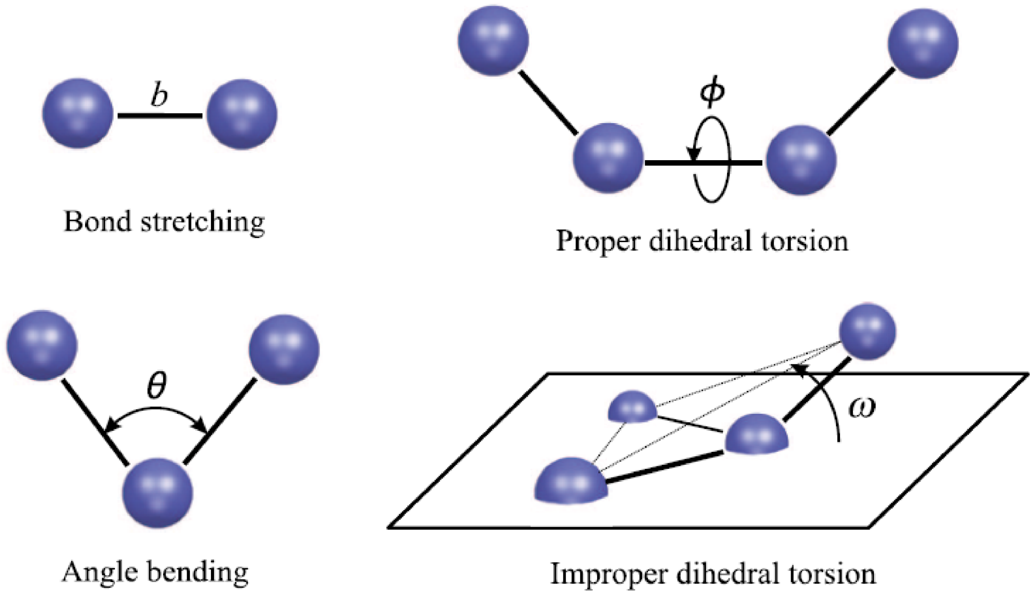


Figure 3.1: Schematic illustration of the bonded terms in the CHARMM force field (adapted from [126]).

3.2.3 Time Evolution

In MD simulations, the motion of the atoms over time are described by Newton's equation of motion:

$$m_i \frac{\partial^2 \mathbf{r}_i}{\partial t^2} = -\frac{\partial U}{\partial r_i}\,, \tag{3.8}$$

where m_i is the mass and r_i the position of atom i, and U is the potential energy of the system. For an efficient numerical time integration of Newton's equation of motion the Verlet algorithm can be used [127].

3.2.4 Simulation Details

MD simulations were performed with NAMD [128] using the CHARMM27 force field for proteins [125], the C35 force field for carbohydrates [129, 130], and the TIP3P model for water [131]. Periodic boundary conditions were used. The long-range electrostatic interactions were computed using the particle mesh Ewald method [132], for which the reciprocal sum was computed with sixth-order interpolation and 1.5 Å Fourier grid spacing. Van der Waals interactions were computed with a smooth switching function between 8 Å and the cutoff value of 10 Å.

The MD systems were first energy-minimized for 5000 steps using the conjugate gradient algorithm and then gradually heated up from 50 K to 300 K over 40 ps. During the minimization and heating phases, harmonic restraints with a force constant of 5 kcal/(mol $\cdot$ Å^2) were applied to the CBM backbone atoms, linker backbone atoms, CD backbone atoms, and the cellulose C_1, C_2, C_3, C_4, C_5, and C_6 atoms. The heated systems then underwent four equilibration steps. In the first 200 ps equilibration run, all solute atoms were fixed, to relax the solvent. In the remaining three equilibration runs, each 200 ps long, all the harmonic restraints were gradually lifted from 5 to 1 kcal/(mol $\cdot$ Å^2). Langevin dynamics was used to maintain constant temperature at 300 K and the Nosé-Hoover Langevin piston [133] with a decay period of 500 fs was used for maintaining constant pressure at 1.01325 bar. After the equilibration, NPT production runs were performed for 60 to 250 ns. A time step of 1 fs for numerical integration of the equations of motion was used. Coordinates were saved every 1 ps.

3.3 Comparison of BD and MD

In this thesis the two different simulation techniques BD and MD are used which are suited to understand the Cel7A-cellulose interaction occurring on different time and length scales (Section 4.3). Both have their advantages and disadvantages. In the following, a short summary of the differences is presented:

- **Solvent:** Virtually all in nature existing biological systems require an aqueous milieu for living [134]. For biomolecules solvent is crucial, it stabilize biological active conformations, and by actively participating in biological processes. Hence, in computer simulations solvent effect have to be included. Here explicit solvent MD simulations are performed, in which each solvent molecule is explicitly modeled. Simulating each solvent molecule explicitly is too computationally intensive for use in simulations covering long time scales. Therefore, in the implicit solvent BD simulations, the Cel7A and cellulose fiber diffuse in a continuum solvent, which reduces the total number of atoms of the system. The implicit solvent model can represent most solvent effects in a quick but nevertheless efficient way. Additionally, in implicit solvent models there is an instantaneous dielectric response of the biomolecules to the solvent, there is no need for lengthy equilibration as in the case of explicitly modeled water molecules.

- **Internal flexibility:** The BD and the MD simulations are both simulated in atomic detail, each single Cel7A and cellulose fiber atom can interact with each other. In the BD simulations, in contrast to MD simulations, the Cel7A and the cellulose fiber are modeled as rigid bodies without internal flexibility. This elimination of nonessential degrees of freedom for the CBM-fiber and Cel7A-fiber encounter process allows the increase of the integration time step (Section 4.1).

- **Bond breaking:** The breaking up of cellulose chains into the individual sugar units requires the breaking of chemical bonds. This process can not be simulated using MD or BD simulations and require QM simulations [135–138] (Section 4.1). Therefore, in this thesis bond breaking events are not studied.

In summary, which simulation method is more suited depends on the biological question asked.

Go where the mess is, there is the action.
Here you can be creative!

Prof. J. Hüfner, Heidelberg University

Chapter 4

Multiscale and Enhanced Sampling

Lets assume we would be living in a world with infinite fast computers and an infinite amount of computational resources. Lets further assume that talented experimentalist have determined with their sophisticated experimental setup a high-resolution structure of a biomolecule [139]. With the first click of the mouse we would get all the required trajectories simulated in the twinkling of an eye, without any significant approximations. After the second click of the mouse, all the simulation trajectories would be analyzed. If we would be living today in such a futuristic world the work presented in this chapter would not be required.

The large number of atoms (approx. 350,000 atoms) and their interactions with each other make atomically detailed simulations computationally quite expensive. All current computational approaches introduce approximations in the simulational setup to ease this problem (Chapter 3). Even if we assume Moore's law [140] would continue to hold, we will not even come in the next decades close to such an ideal state for the Cel7A-cellulose system. In this chapter first, the *time and length scales of biomolecules* is introduced and in the second part put in relationship to *Moore's law*. Third, the *enhanced multiscale approach to combine BD and MD simulations* is described. Fourth, the approach to *cover the configuration space* of Cel7A-cellulose fiber is presented. Fifth, the *course of dimensionality* and the *dimension reduction* strategy is explained. Seventh, the *Markov state model* (MSM) approach is introduced. Eight, the *density map* analysis is described. Which provides a thermodynamic description of biomolecular interaction, in contrast to the MSM analysis which rather gives a kinetic description. In the last section, some details on the *number crunching* and *big data*, which was required to perform this research, is presented.

> There is nothing insignificant in the world. It all depends on the point of view.
>
> *Johann Wolfgang von Goethe 1749-1832*

4.1 Time and Length Scales of Biomolecules

The biological functions inside and between different biomolecules are governed by the interactions of individual atoms and groups of atoms. These interactions occur at multiple time scales, spanning more then 15 orders of magnitudes between them. The range starts from femto-seconds (10^{-15} s) and can go beyond the second time range (Figure 4.1). Computer simulation methods have become a powerful tool to study problems in biophysics [141]. The accessible time and length scale coverable for each simulation method are different. The more detailed the simulation technique operates, generally the smaller is the accessibility of long time and length scales [141]. The fastest motion of the modeled system determines the length of the integration time step, which directly determines the amount of simulation steps required to reach a certain simulation length [142].

Using quantum mechanics (QM) in principle a very high level of accuracy is obtainable. In QM simulations the fast motion of electrons are taken into account, they are the most restricted in terms of the maximal time and length scales. However the time to compute processes on the long time and length scale of biomolecular interactions using a fully QM description is not possible.

In classical MD the electronic distribution are approximated in a rather classical coarse-grained fashion, for example by putting fixed partial charges on the interaction sites of the biomolecules [141]. In MD the time step of integration is dominated by events like rotational motions and intermolecular vibrations of side chain groups, which are an order of magnitude slower then the electron motion.

In BD one is in general not interested in the detailed description of the solvent. The effect of the solvent is modeled using an implicit solvent model. The second approximation is, that the internal flexibility of the biomolecules is neglected, by treating each biomolecule as a rigid full atomic body. The global translation and rotation of the biomolecules are the fastest motions in the system, which are an order of magnitude slower compared to classical MD.

QM has among the above three methods the least level of approximation and is better suited to simulate fast electronic motions and bond making and breaking events. MD on the other hand gives a local picture of the biomolecule interaction, e.g. like side chain motion, and BD gives a global picture of the interaction, e.g. like the encounter process of biomolecules. The integration time step together with the fact, that the accessible computational resources are limited determines directly the total simulation time of each simulation method and the typical range of observable biomolecular motion. Observing an event a single or even a few time is in general not sufficient to derive a meaningful conclusion for a biomolecule. To determine if the observation was a random event or a statistically relevant feature, even longer simulation times are required, which make the situation even more problematic.

With a single simulation methodology, it is at the moment extremely difficult to achieve the required precision and cover the relevant motion range of the Cel7A-cellulose interaction. Therefore, with the available supercomputers multiple methods together with a framework to combine the simulation results in a meaningful way are indispensable for a continuous analysis of the Cel7A-cellulose interaction, over the entire relevant time and length range.

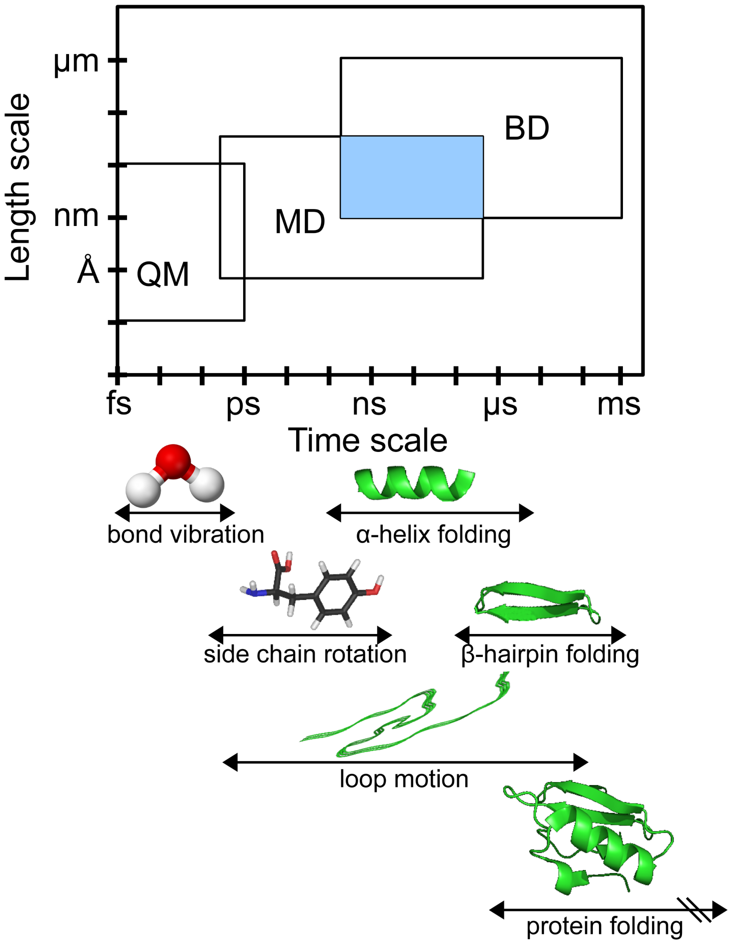


Figure 4.1: Diagram of time and length scales for simulation methods including quantum mechanics (QM), molecular dynamics (MD), and Brownian dynamics (BD).

4.2 Moore’s Law in Molecular Biology

At the moment it is extremely challenging, or even impossible, to simulate the Cel7A-cellulose complex using classical MD on a ms time scale. Moore’s law applied to molecular biology can give an indicator at which time in the future this will be possible.

In classical MD simulation Newtons’s equation of motion are used to describe the dynamics of biomolecules and to enhance our understanding of atomic processes in living systems. The basic theoretical background was already set in the 17th and the early 20th century. The application for relevant biological systems was hindered by the required extensive numerical calculations. The possible applications strongly depends on the computational resources at hand. The breakthrough was achieved with the upcoming of the first supercomputers [142–145]:

- 1687: Newton published his *laws of motion* for classical physics.
- 1860: Pasteur and Hofmann introduced the *ball and stick model* to decribe the molecular structure of chemical compounds.
- In the first half of the 20th century the concept of the *force field* originated to decribe the forces acting between the atom pairs in biomolecules.
- 1953: Scientist from Los Alamos published their study “Equation of State Calculation by fast Computing Machines” [146] which laid the groundwork for *Monte Carlo* and *MD* simulations. The calculations were performed on the MANIAC supercomputer in Los Alamos, which was the birth of *computational physics*.
- 1957: First molecular dynamics simulations were carried out [147].
- 1964: Newton’s Mechanics was combined with the biomolecular force field approach for argon. This combination would be later known as the *molecular mechanics method* [148].
- 1971: First simulation of water by Rahman and Tillinger (216 molecules) [149].
- 1977: First simulation of the BPTI protein in the absence of solvent molecules (in vacuum) (≈500 atoms, ≈10 ps) [150].
- 1982: First simulation of the BPTI protein in explicit water. The explicit modeling of the the solvent molecules makes the simulations computational more expensive (≈3,100 atoms, ≈25 ps) [151].

- 1992: Simulation of the HIV-1 protease in explicit solvent using a CRAY YMP computer (≈23,000 atoms, ≈40 ps) [152].
- 1998: First report of a 1 ms MD simulation of the villin headpiece subdomain (≈3,100 atoms, ≈1 ms) [153].
- 2002: Explicit solvent simulation of FN-III (126,000 atoms, ≈12 ns) [154].
- 2004: Explicit solvent simulation of the DOPC lipid bilayer (420,000 atoms, ≈3.5 ns) [155].
- 2008: Explicit solvent simulation of the WW domain on an x86 cluster (≈10 μs) [156].
- 2008: The special-purpose MD parallel supercomputer *Anton* becomes operational. It can simulate over 17,000 ns per day for a protein-water system consisting of 23,558 atoms[1].
- 2010: The 8.8 PFlop/s distributed-computing project Folding@home achieves the aggregate ensemble simulation time scale of 1.5 ms [157].
- 2010: Explicit solvent simulation of the folded BPTI and the WW domain on Anton (≈1 ms) [158].

Over the last decades a dramatic progress in MD can be observed, the length of the trajectories and the amount of atoms have significantly increased. The progress amongst others can be described by Moore's law [159], the observation that over last 50 years the number of transitions on integrated circuits doubles approx. every 20 months [159–161]. From this follows a similar growth in computing power. The number of atoms and the time scale accessibility of biomolecules simulation follow a similar quasi Moore's trend [143, 144]. In the milestone MD simulation presented in Figure 4.2, either very large biomolecules are simulated for very short time scales or very small biomolecules are simulated for very long time scales. The application of the concept of Moore's law to biological systems indicate that every 10 years the number of atoms increase by approx. one and a half order of magnitude (Figure 4.2, top) and that the accessible time scale increase by approx. three orders of magnitude (Figure 4.2, bottom).

The combination of both trends gives an indicator at which time in the future it will be possible to perform MD simulations for the Cel7A-cellulose complex on the ms

[1] http://www.nrbsc.org/anton_rfp/

time scale without an enhanced multiscale approach. The Cel7A-cellulose system has ≈350,000 atoms. In 2004 a 3.5 ns MD simulation for a biomolecule consisting of 420,000 atoms was performed [155]. Using Moore's law a decade later, in 2014, the μs time scale will be accessible for the Cel7A-cellulose system. After another decade, in 2024, the ms time scale will become accessible. This shows clearly, that enhanced and multiscale approaches are absolutely essential to study with todays supercomputers molecular process on the ms time scale for the Cel7A-cellulose system.

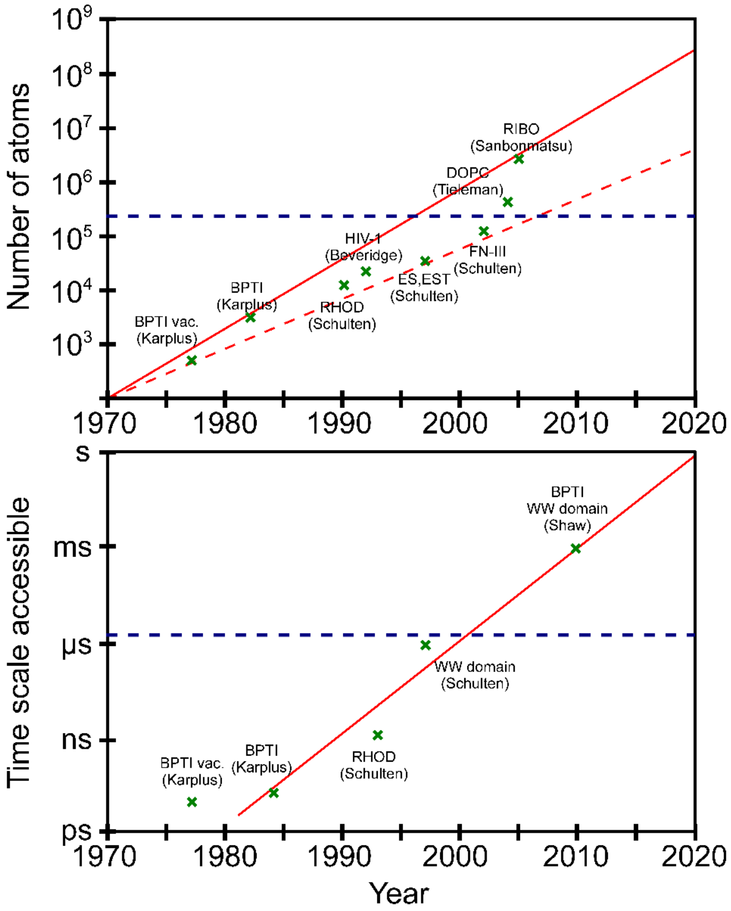


Figure 4.2: (Top) Increase in MD simulation system size with respect to year simulated. All simulations include explicit solvent, except "BPTI vac.". Solid red curve, Moore's law doubling every 28.2 months. Dashed red curve, Moore's law doubling every 39.6 months. Blue curve, MD simulation with largest number of atoms used in this thesis. (Bottom) Growth in time scale accessible to MD simulations of proteins. Red curve, Moore's law doubling every 12 months. Blue curve, sum of simulation time of all MD trajectories is 6.2 μs (adapted from [143, 144]).

4.3 Enhanced Multiscale BD and MD Simulations

Previous simulation studies have provided vital clues at individual time and length scales, however, the detailed understanding of the entire Cel7A-cellulose dynamics requires informations over a broad range of the time and length scale. To access the entire range a single simulation method is not sufficient (Section 4.1). Even with todays powerful computers, extending the time scale of MD to μs or longer for systems of a few hundred thousand atoms or more remains a significant challenge, although μs-ms are time scales of particular biological relevance. For example, with an estimated translational diffusion constant of 16 Å^2/ns, it takes at least hundreds of ns for the CBM to orbit once a cellulose fiber with a radius of 40 Å. A minimum total simulation time of several μs would be required for obtaining statistically reliable results (Figure 4.3). It is essential to use a multiscale modeling approach, to cover the entire range on an atomic level. The use of multiple simulation methods, like this MD-BD approach, brings the challenge how to transfer and combine the informations obtained between the BD and MD level. The difficulty for the Cel7A-cellulose system can be solved by looking at the different steps involved in the deconstruction of cellulosic biomass by Cel7A (Figure 4.4):

(S1) initial encounter of the free floating Cel7a with cellulose,

(S2) diffusion of the CBM on the cellulose fiber surface,

(S3) binding of the CBM to cellulose,

(S4) initial threading of cellulose chain into the catalytic tunnel located in the CD,

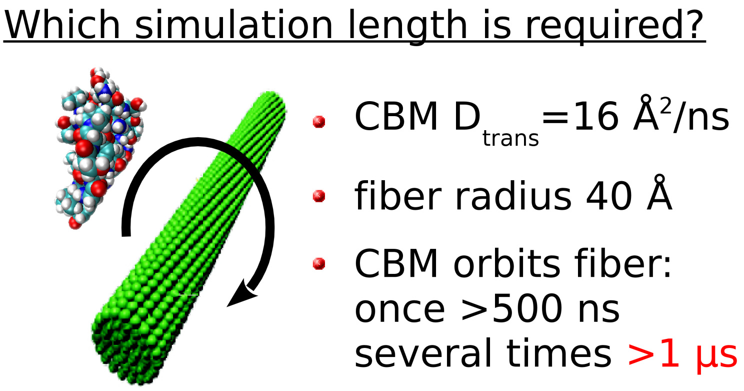
(S5) hydrolysis of cellulose chain, and

(S6) processive threading of the next cellubiose unit.

To study the step (S1) primarily BD, and for (S2)-(S4) MD simulations are required. The steps (S5)-(S6) involve making and breaking of chemical bonds, which require QM simulations and which are outside the scope of this study.

The ansatz used here will be four-fold. Firstly, all-atom, explicit solvent MD simulations will be used to derive and parameterize coarse grain models for Cel7A and cellulose (Chapter 5 and Chapter 6). Secondly, to study the global aspects of the Cel7A-cellulose interactions (S1) all-atomic rigid body implicit solvent BD simulations will be used (Chapter 5). Thirdly, to extend BD to study local interactions (S2 and S3) at the other end of the spectra, again all-atom explicit solvent MD simulations are performed (Chapter 5). Fourthly, with the information from the first three steps it will be possible to parameterize spring-models for the Cel7A linker, allowing a transformation from the all-atom linker to a more coarse-grained spring representation, hereby permitting to derive and simulate optimized artificial cellulase enzymes which currently do not exist in nature and first have to be genetically engineered (Chapter 6).

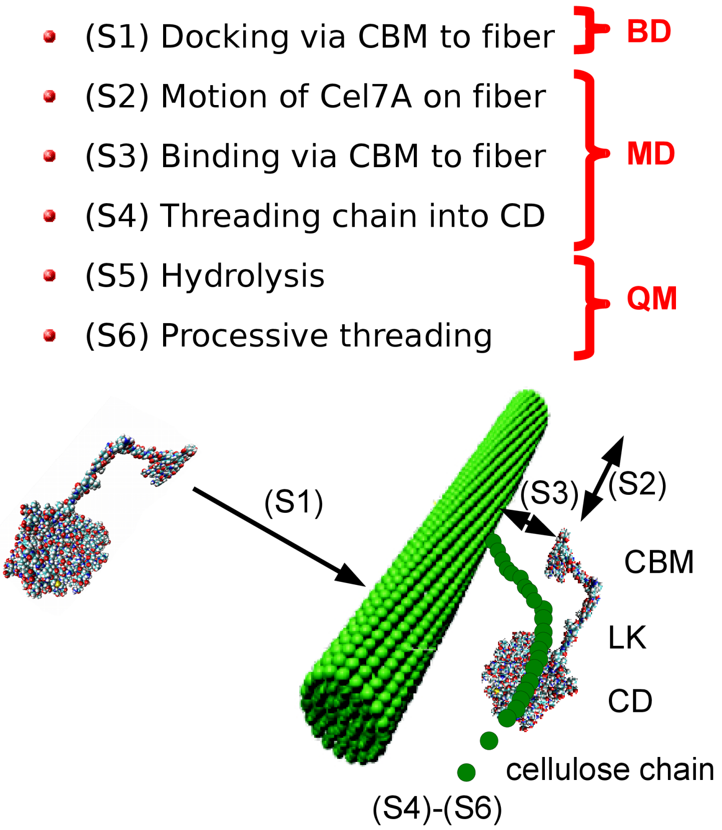
The main goal of this work is to cover each step of the Cel7A-cellulose interaction with the required accuracy on an atomic level. Finally, the described ansatz allows us with the available computational resources to understand the role of the CBM and the linker peptide in the interaction with the cellulose fiber.



Which simulation length is required?

- CBM D_{trans}=16 Å^2/ns
- fiber radius 40 Å
- CBM orbits fiber:
  once >500 ns
  several times >1 µs

Figure 4.3: With an estimated translational diffusion constant of 16 Å^2/ns, it takes more then 500 ns for the CBM to orbit once a cellulose fiber with a radius of 40 Å, and at least several μs o obtain statistically reliable results.



- (S1) Docking via CBM to fiber
- (S2) Motion of Cel7A on fiber
- (S3) Binding via CBM to fiber
- (S4) Threading chain into CD
- (S5) Hydrolysis
- (S6) Processive threading

Figure 4.4: The deconstruction of cellulosic biomass by Cel7A can be subdivided into the steps (S1) to (S6). Different computer simulation techniques including quantum mechanics (QM), molecular dynamics (MD), and Brownian dynamics (BD) are required for each step.

The wise man looks into space and he knows there is no limit to dimensions.

Master Zhuang 369 BC-286 BC

4.4 Cover Configuration Space

To cover the relevant CBM-fiber interaction space, BD simulations were started from 2,800 different configurations. To generate these configurations, the CBM was rotated randomly and its center placed in a random position in one of three shells around the fiber. As displayed in the yz plane (Figure 4.5, top left), the first shell was defined by a ring with a radius of r_R=60 Å from the fiber center; the second shell was an octagon with a distance of r_O=21 Å from the fiber surface C_1 atoms; and the third shell was a ring segment only above the hydrophobic fiber face (1, 0, 0) and the hydrophilic fiber face (0, 1, 0) with a radius of r_S=40 Å from the fiber center, which was intended to examine the association/dissociation kinetics of the CBM from these two fiber faces. 300 BD simulations were started from the first r_R shell, 500 from the second r_O shell, and 1,000 from the third r_S shell (Figure 4.5, bottom). With the constraint of available computational resources, the combination of these three shells was expected to produce less biased results compared to a single shell. The comparison of the binding probability results obtained from different shells also provided a measure of convergence (i.e., statistical errors) of the BD simulations.

Experimental studies found that the hydrophobic fiber faces are the interaction hotspots of Cel7A [40–42]. To study the effect of the linker on the Cel7A-fiber interaction I focused on the hydrophobic fiber faces. In the BD starting structures the center of Cel7A is placed on a 130 Å circle segment above the hydrophobic fiber face (1, 0, 0) (Figure 4.6).

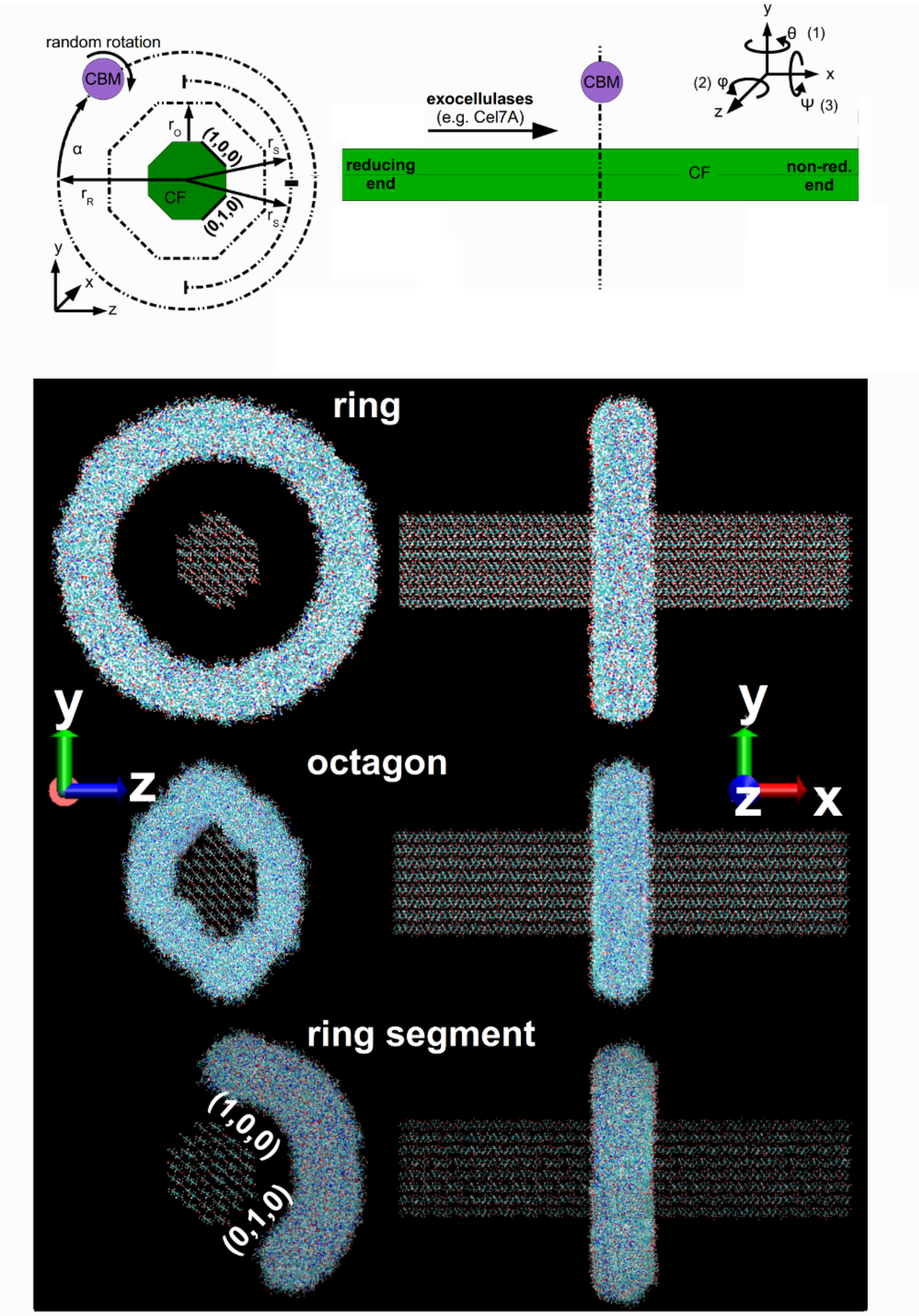


Figure 4.5: (Top) To ensure the BD simulations to cover all relevant CBM-cellulose configurations, the CBM was initially placed at either a distance r_R or r_S from the center of the cellulose fiber or at a distance rO from the fiber surface. Starting configurations of the BD simulations of the CBM from different shells are shown in an yz plane view (bottom left) and an xy plan view (bottom right).

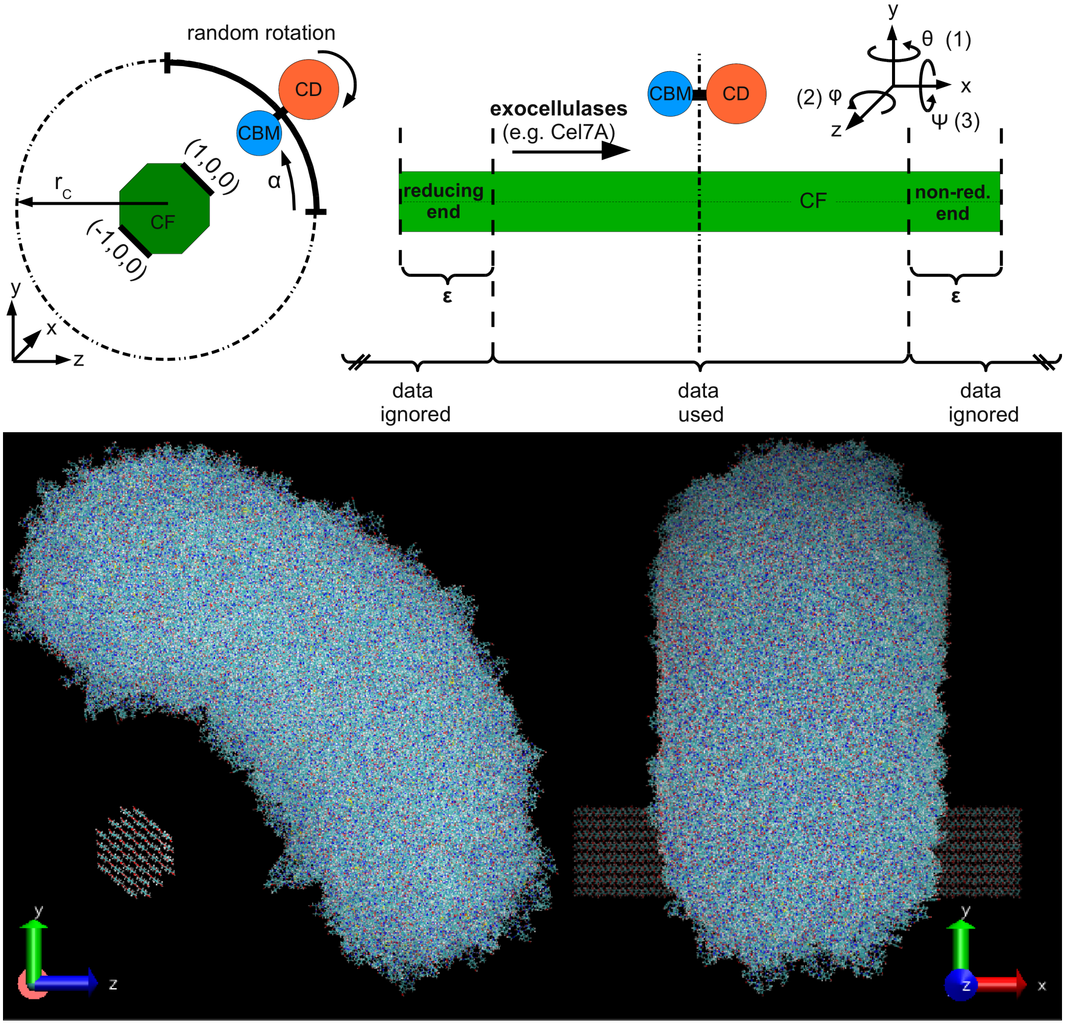


Figure 4.6: (Top) Sketch of the Cel7A-fiber complex. To ensure that the BD simulation trajectories cover the relevant configuration space, Cel7A is placed in a distance $r_C = 130\,\text{Å}$ from the center of the cellulose fiber on a 90° circle segment above the (1, 0, 0) hydrophobic fiber face. This ensures that the starting configurations of the trajectories are all statistically independent. The direction of the rotation axis of Cel7A and the rotation angle α is chosen randomly. (Bottom) In order to visualize the different starting structures of the BD simulation trajectories the starting structures are super imposed. (Bottom left) yz plane view and (bottom right) xy view.

Everything should be made as simple as possible, but not simpler.

Albert Einstein 1879-1955

4.5 Dimension Reduction

Atomic detailed simulations represent inherently high dimensional *big data* sets. The complex bimolecular interactions imply that only a small number of dimensions may be relevant for the biological function. For an n atomic biomolecule data-sets can be $3n$ dimensional. For the Cel7A-cellulose system with approx. n =350,000 atoms this gives over one million dimensions! Analyzing and organizing data in such high-dimensional spaces various negative phenomena, often referred to as *course of dimensionality* [162, 163] can be observed which do not occur in low dimensional, e.g. three dimensional physical space. One of the main problems is, that the increasing dimensionality, the volume of the space increases so fast that the data distribution inside the space appears sparse and dissimilar, which prevents some algorithms to be numerically efficient (Figure 4.7). Dimension reduction methods map each $3n$ dimensional bimolecular conformation to a data point on a lower dimensional manifold. They reduce the degrees of freedom of the system and improve the numerical stability. The complexity of biomolecules makes such a general mapping non trivial.

To reduce the number of degrees of freedom needed in the analysis of the BD simulations, following dimension reduction strategy was applied for the CBM-cellulose and Cel7A-cellulose complex (Figure 4.8). In an n atomic biomolecule, each atom can move in all three space directions x, y, and z, which results in $3n$ degrees of freedom. In the BD simulation, the internal flexibility of the biomolecules is neglected. The CBM and the celluose fiber can only perform a translational and rotational motion, this reduces the degrees of freedom from $3n$ to 12. Prior to analysis all trajectories are transformed into a reference system where the cellulose fiber was fixed, it has no translational or rotational motion. This corresponds to the coordinate system where the physical observer is placed on the cellulose fiber and watches the motion of the CBM or CD, respectively. This step reduces dimensionality of the system from 12 to 6 degrees of freedom (translation and rotation of the CBM or CD, respectively). The center of the cellulose fiber was placed at the origin and following orientation of the coordinate system was used. The chain direction of the fiber is oriented along the x-axis of the coordinate system, with the reducing end of the fiber pointing to the negative direction and the non-reducing end to the positive direction. The other two shorter dimensions were oriented parallel to the y and z axis of the coordinate system (Figure 4.9). The rotational angle around the x axis was denoted as ψ, around the y axis as θ and around the z axis as ϕ. The fiber is translated such that its center lies at the origin of the coordinate system. An infinite long fiber is homogeneous along the x axis. For the CBM binding with the different fiber faces, the x dimension degree of freedom was integrated out. For the BD simulations a fiber with an finite length along the x axis is used. To avoid the potential artifact arising from the CBM interacting with the fiber ends, only those simulation frames in which the center of the CBM or CD, respectively, was located within 30 Å from the fiber ends were excluded from the analysis (Figure 4.9). Only the three rotational orientations and the y and z component of the CBM or CD position, respectively, are required to describe the spatial relationship of the CBM with respect to the cellulose fiber. This reduces the total degrees of freedom of the CBM-cellulose and CD-cellulose system from over one million to five.

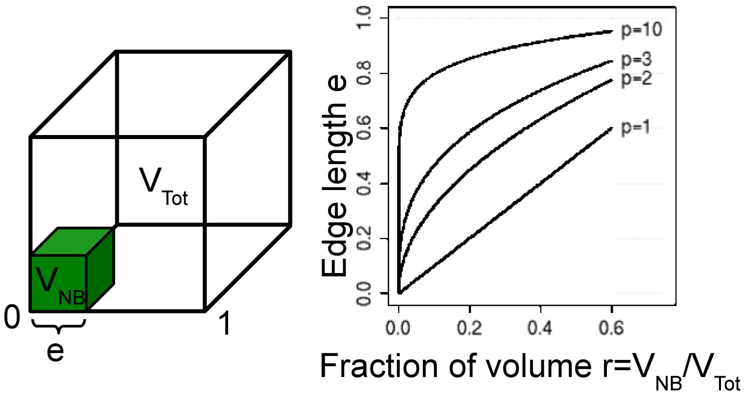


Figure 4.7: To visualize the curse of dimensionality, lets assume the data is uniformly distributed in a unit hypercube with volume $V_{Tot} = 1$. Lets define a neighbor volume $V_{Nb} = e^p$, with edge length e, and number of dimensions p. The fraction of unit data volume is given as $r = V_{Nb}/V_{Tot} = e^p$. As the dimension p increases the distance to neighbor, data points increase. For p=10 dimensions to capture $r = 1\,\%$ of the data, the neighbor volume must cover $e = 63\,\%$ of the range of each input variable (95 % for p=100) (adapted from lecture "Introduction to Machine Learning, Pattern Recognition and Statistical Data Modeling" by Dr. Coryn Bailer-Jones).

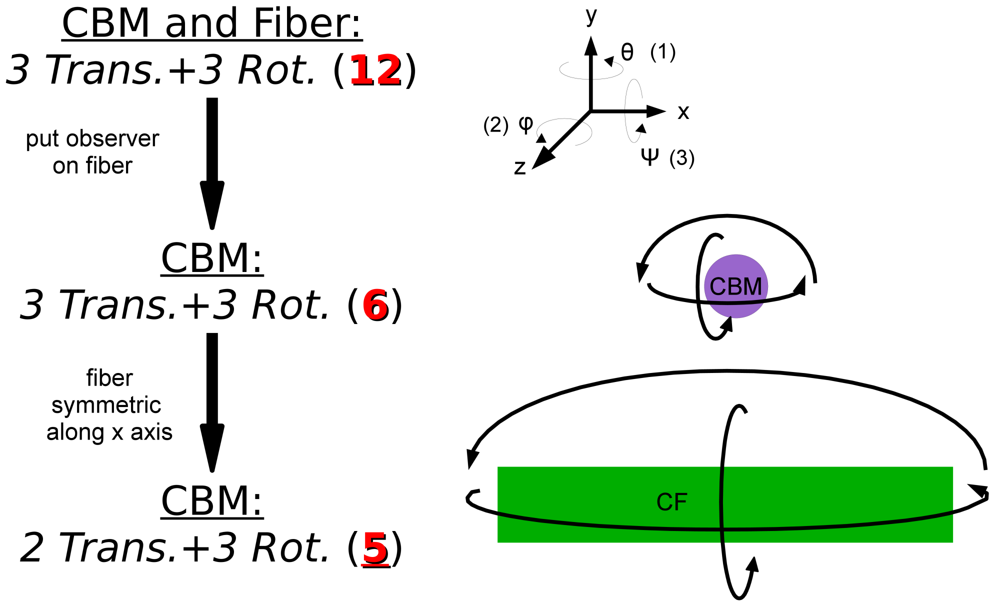


Figure 4.8: To facilitate the construction of the MSM the dimensionality of the system has to be reduced. By ignoring the internal flexibility of the atoms the degrees of freedom can be reduced from over one million to 12. Using the protocol above it can be further reduced from 12 to 5.

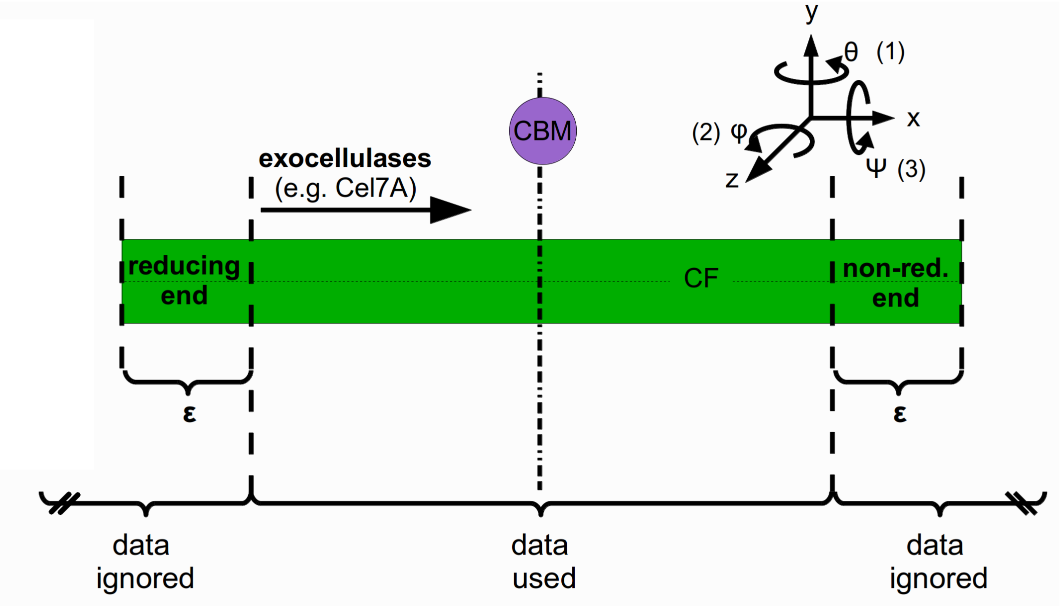


Figure 4.9: Based on the limited computational resources it is impossible to simulate an infinite long cellulose fiber. To still be able to model an infinite long cellulose fiber, the dynamic spectra is analyzed on the Markov lag time scale τ and only the dynamics around the fiber center is included, ignoring the dynamics within ϵ =30 Å from the fiber ends.

Memory is deceptive because it is colored by today's events.

Albert Einstein 1879-1955

4.6 Markov State Models

Simulations in the range of ms are required to study the Cel7A-cellulose fiber encounter process with statistical accuracy. Even, if BD instead of MD simulations are performed, with the given computational resources it is very challenging to sample the relevant configuration space with a single BD trajectories of this length. To still tackle this problem with today's supercomputers a large set of independent BD simulations are computed in parallel, and the statistical information is combined using a Markov state model (MSM) framework [164–172]. The MSM description provides a mathematically rigorous way to combine the statistical information from multiple independent simulations and describe the dynamics of this system. From the simulation trajectories a transition matrix **T** can be calculated, which describes the interaction between different biomolecules. The MSM approach allows to apply a set of interesting properties from Markov theory to a given biological system (Figure 4.10). The main assumption is, that the transition between bimolecular conformations are Markovian (memoryless). This means the transition depends only on the current conformation and not the previous conformations. Such an assumption is not uncommon, e.g. Newton's laws of motion, which describe classical mechanics, makes a similar assumption. The future position and velocity of a particle only depends on the current position and velocity (Figure 4.11). The lag time τ denotes the time scale after which the transition between the two conformations becomes Markovian. The implied timescale test was used to estimate τ [173, 174]. The cellulase enzyme can interact with the cellulose fiber. During the simulation different parts of the configuration space of the biomolecule are visited. Defining a MSM for a given biomolecule can be quite a challenge, following steps are required:

1. The configuration space of the biomolecule has to be discretized to find similar conformations (microstates).

2. The smallest time scale τ, for which the transitions between the conformations are Markovian has to be estimated.

3. The transition probabilities between this conformation has to be calculated on the time scale τ.

The conformations of a biomolecule together with the transition probabilities between conformations define a MSM. To demonstrate this, lets look at a simple toy example with the three conformations 1, 2, and 3 (Figure 4.12). Lets assume the Markov time scale τ is given and that the biomolecule is at time t in conformation 1, at time $t + \tau$

it can either stay in conformation 1 with the probability p_{11} or jump to conformation 2 or 3, with probability p_{12} or p_{13}, respectively. Similar transitions can proceed from conformation 2 or 3. In total nine different probabilities p_{11}, $p_{12}, \ldots, p_{33}$ can be conceived. A transition matrix $\mathbf{T}$ can be written for these probabilities, which describes the interaction of the Cel7A with the cellulose fiber for time scales $> \tau$. To construct a MSM for the CBM-cellulose system following steps were performed. Given a set of microstates, the transition count matrix was constructed as

$$\mathbf{C}(\tau) = (c_{ij})_{1 \leq i,j \leq n} \ . \tag{4.1}$$

where c_{ij} counts the number of times the CBM visits state i at time t and state j at time $t + \tau$, for all times t. A reasonable (maximum likelihood) estimate for the true transition probability t_{ij} between states i and j was then given by

$$\forall_{1 \leq i,j \leq n} \quad p_{ij} = c_{ij} / \sum_{j} c_{ij} \tag{4.2}$$

The transition probabilities form a transition matrix

$$\mathbf{T}(\tau) = (p_{ij})_{1 \leq i,j \leq n} \ . \tag{4.3}$$

The transition matrix, together with the discretized microstates, defines the Markov state model, from which various properties of the CBM-fiber interaction can be calculated (Figure 4.13). For an ergodic system, $\mathbf{T}(\tau)$ has only a single left eigenvector with eigenvalue 1:

$$\pi = \pi \cdot \mathbf{T}(\tau) \, . \tag{4.4}$$

Normalizing this eigenvector leads to the stationary distribution $\pi = (\pi_1, \pi_2, \ldots, \pi_n)$, where each element π_i gives the probability of microstate i (finding the CBM in the fiber region i). The diagonal element p_{ii} of $\mathbf{T}(\tau)$ gives the probability of the CBM being trapped in the fiber region i once it reaches there, providing a measure of how sticky the fiber region i is. The corresponding mean exit time associated with this probability is

$$t_{exit,i} = -\tau / ln(p_{ii}) \, , \tag{4.5}$$

which provides the average time for which the CBM will stay bound to the fiber region i before escaping. To estimate statistical uncertainties of the properties derived from the MSM model, the Bayesian statistics approach [167, 168] was used. To make the MSM calculation tractable, a coarse discretization was used. Here, the yz space was partitioned into a total of n=100 fiber regions (microstates), each covering a circular segment of 18°and a radius of 10 Å. The yz space between the radii of 20 and 30 Å from the fiber center forms the innermost ring while that with the radius of $>$ 60Å from the fiber center constitutes the outermost. The data analysis was performed using GNU R [175].

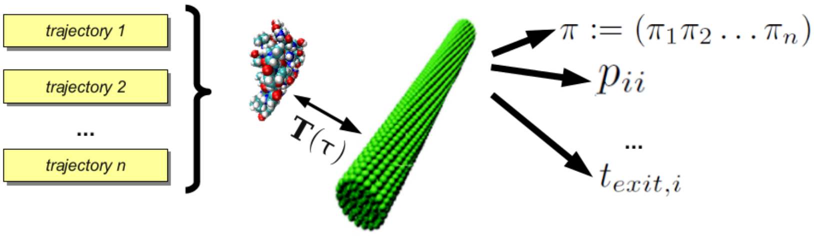
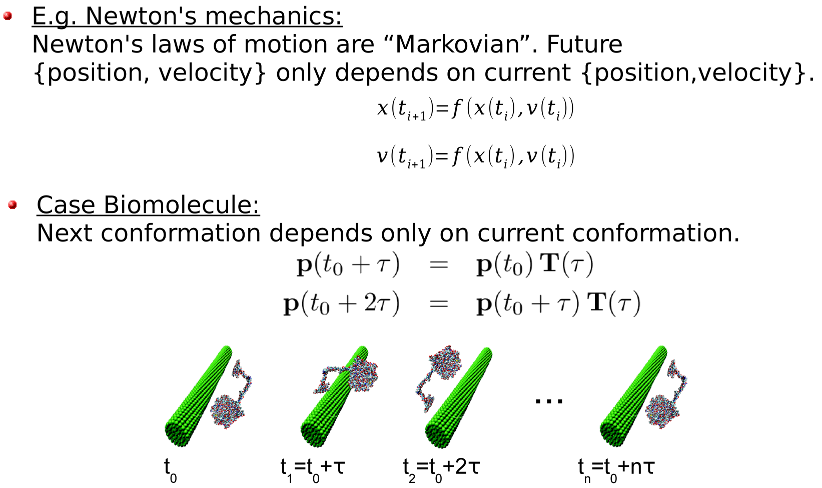


Figure 4.10: The MSM frame work allows it to combine the statistical information from multiple independent simulations. The transition matrix **T** contains the information on interaction of the CBM with the cellulose fiber. Once a MSM is defined, various concepts from Markov chain theory can be applied to the biological system.



- E.g. Newton's mechanics:
  Newton's laws of motion are "Markovian". Future {position, velocity} only depends on current {position,velocity}.

$$x(t_{i+1}) = f(x(t_i), v(t_i))$$

$$v(t_{i+1}) = f(x(t_i), v(t_i))$$

- Case Biomolecule:
  Next conformation depends only on current conformation.

$$\begin{aligned} \mathbf{p}(t_0+\tau) &= \mathbf{p}(t_0)\,\mathbf{T}(\tau) \\ \mathbf{p}(t_0+2\tau) &= \mathbf{p}(t_0+\tau)\,\mathbf{T}(\tau) \end{aligned}$$

Figure 4.11: In the MSM context it is assumed, that the transition between biomolecule conformations are memoryless, which means the future conformation only depends on the current conformation and not the previous conformations. Such an assumption is not uncommon, e.g. Newton's laws of motion, which describe classical mechanics, makes a similar assumption.

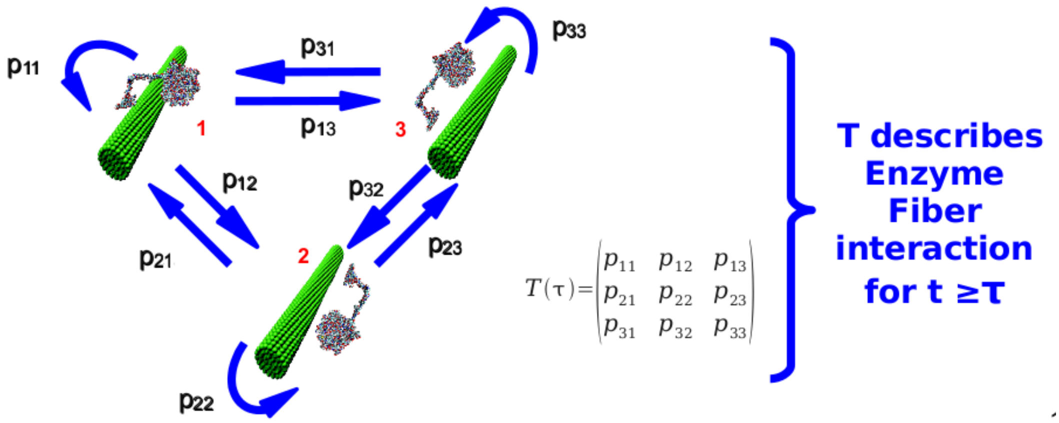
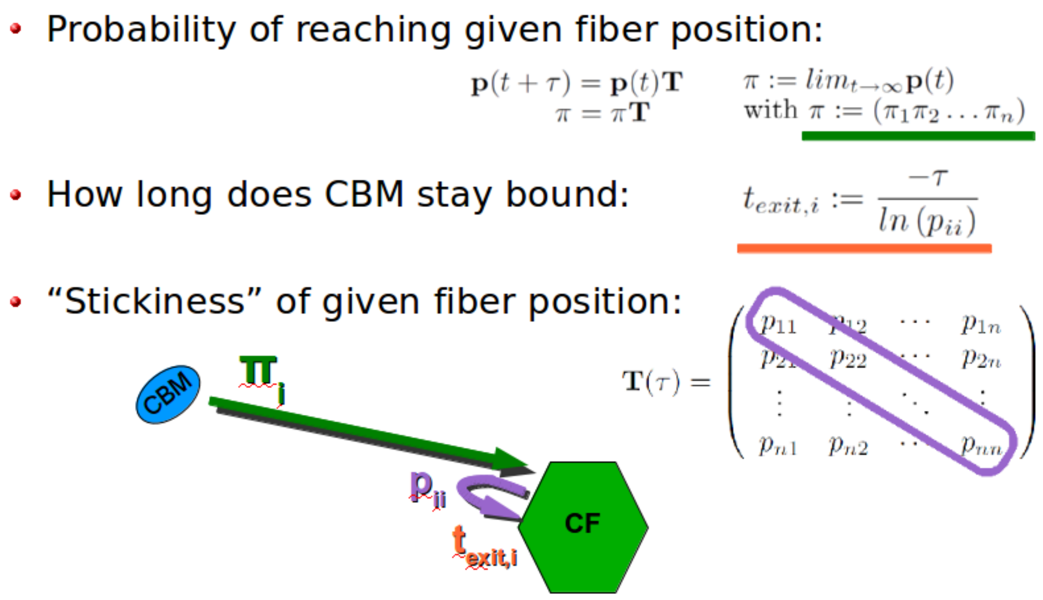

p11
p31
p13
p33
1
3
p12
p32
p21
2
p23
p22
$T(\tau)=\begin{vmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{vmatrix}$
T describes Enzyme Fiber interaction for t ≥τ

Figure 4.12: Simple three conformation toy example of a MSM model. A MSM is defined as a set of conformations, the smallest Markov time scale τ, and the transition probabilities between the conformations, which are summarized in a transition matrix $\mathbf{T}$.



- Probability of reaching given fiber position:

$$\mathbf{p}(t+\tau) = \mathbf{p}(t)\mathbf{T} \qquad \pi := lim_{t\to\infty}\mathbf{p}(t)$$
$$\pi = \pi\mathbf{T} \qquad \text{with } \pi := (\pi_1\pi_2\ldots\pi_n)$$

- How long does CBM stay bound:

$$t_{exit,i} := \frac{-\tau}{ln\,(p_{ii})}$$

- “Stickiness” of given fiber position:

$$\mathbf{T}(\tau) = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{pmatrix}$$

CBM
π_i
p_{ii}
$t_{exit,i}$
CF

Figure 4.13: From a MSM description, various properties of the CBM-fiber interaction can be calculated, like the stationary distribution π, the mean exit time $t_{exit,i}$, and the transition probabilities p_{ii}.

4.7 Density Map

To determine the probability of each fiber face being visited by the CBM a density map was calculated from the BD trajectories. The y and z dimension that range from -60 Å to 60 Å were each discretized into 120 bins, resulting in a total of 14400 bins covering the yz plane (Figure 4.5). The probability $p(x, y)$ for each bin was obtained by counting the number of times $h(y, z)$ each bin was visited by the center of the CBM, followed by normalizing by the total number of visits to all the bins h_{tot}. $p(y, z)$ was further converted to a relative free energy by computing the negative natural logarithm as $F(x, y) = -k_B \cdot T \cdot ln\,[p(y, z)] + c_0$, where c_0 is a reference constant.

To determine how often a fiber face was visited, the observations of the 50 and 1600 most visited bins close to a specific fiber face are summed up to

$$h_{50} = \sum_{j=1}^{50} h(j) \tag{4.6}$$

and

$$h_{1600} = \sum_{j=1}^{1600} h(j)\,. \tag{4.7}$$

$h(y, z)$ was sorted in decreasing order, the counting index in the sorted set $h(j)$ is denoted as j. The area around the fiber face (1, 0, 0) resp. (-1, 0, 0) hotspot for the h_{50} calculation has 400 bins and is defined as a square with $3\,\text{Å} < y \leq 23\,\text{Å}$, $10\,\text{Å} < z \leq 30\,\text{Å}$ resp. $-23\,\text{Å} < y \leq -3\,\text{Å}$, $-30\,\text{Å} < z \leq -10\,\text{Å}$ (Figure 4.14). For the calculation of h_{1600} a larger square around the hotspot with 1600 bins was chosen, for the fiber face (1, 0, 0) resp. (-1, 0, 0) hotspot the square is defined as $0\,\text{Å} < y \leq 40\,\text{Å}$, $0\,\text{Å} < z \leq 40\,\text{Å}$ resp. $-40\,\text{Å} < y \leq 0\,\text{Å}$, $-40\,\text{Å} < z \leq 0\,\text{Å}$. The choice of h_{50} and h_{1600} allows zoomed in or a zoomed out view of the fiber face hotspot, respectively. In the starting structure of the BD trajectories the Cel7A was placed above the fiber face (1, 0, 0), the fiber face (-1, 0, 0) being the furthest away from this starting position (Figure 4.6 top left). h_{50} and h_{1600} for the fiber face (-1, 0, 0) gives an estimate for the CBM sampling speed of the cellulose fiber.

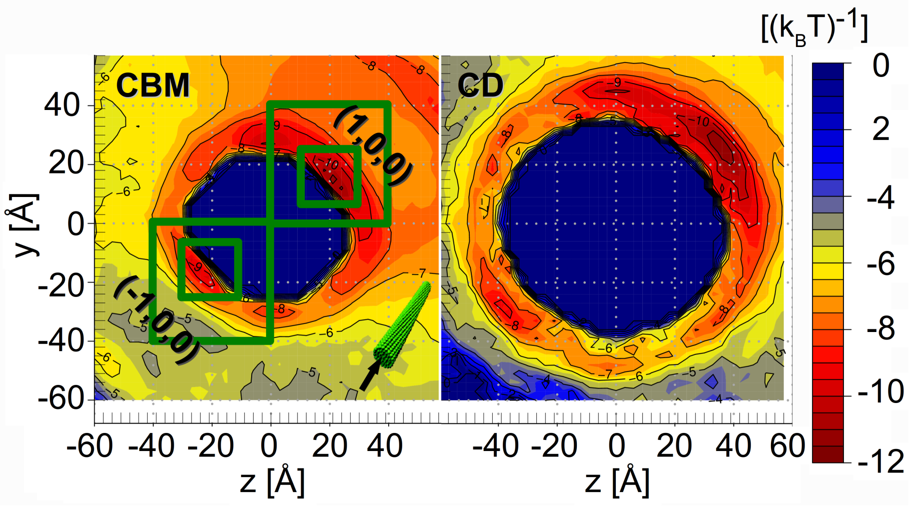


Figure 4.14: BD simulation density map of the CBM (left) and the CD (right). The relative free energy F(y,z) is plotted as a function of y and z coordinates. The smaller and larger green squares around the fiber face (1, 0, 0) and (-1, 0, 0) hotspots visualize the area for the h_{50} resp. h_{1600} calculation.

> The way to think of a supercomputer is as a special-purpose device. Only with these devices can we perform this cutting-edge research.
>
> *Jack Dongarra 1950*

4.8 Number Crunching and Big Data

For this study and possible follow-up research (Section 7.2), simulations were performed and analyzed on supercomputers on two different continents (Table 4.1). The main systems were *Kraken*[2], *Franklin*[3], *Hopper*[4], *Jaguar*[5], and *Titan*[6] in the USA and *bwGRiD*[7] in Germany. The simulated data is stored at the high performance storage systems at the *National Energy Research Scientific Computing Center* (NERSC), the *Oak Ridge National Laboratory* (ORNL), and the *bwgrid-storage* at the *Karlsruhe Insitute of Technology* (KIT). To create, handle, and analyze the data code was written in the programming languages C/C++, Java, Fortran, Perl, and Bash. The computational chemistry packages CHARMM [176], GROMACS [177], and NAMD [128] were used. For the statistical analysis code was written in Matlab/Octave [178–180] and GNU R [175]. An overview of the research in numbers:

- Biological system has approx. 350,000 atoms, full atomic BD and MD simulations were performed
- CBM study (Chapter 5):
  BD: $> 7{,}600$ trajectories; > 29 ms
  MD: 54 trajectories; > 1.2 μs
- Linker study (Chapter 6):
  BD: $> 47{,}000$ trajectories; > 47 ms
  MD: 47 trajectories; > 5.2 μs
- Consumed storage space: > 60 TB
- Total computation time (simulation and analysis): > 7 million CPU hours

If we assume that societies invest more resources on problems they perceive as important and that the consumed CPU hours are a meaningful key indicator, then a rough estimate on the importance of the research presented in this thesis can be calculated. In total approx. 7 million CPU hours were used (Table 4.1) for this and follow-up projects, with a cost of \$0.10 per CPU hour [181], this results in a total economical value of \$700,000. To put this in perspective, this corresponds to 4.2 ‰ of the World Health Organization yearly spending to fight malaria [182].

[2] http://www.nics.tennessee.edu
[3] http://www.nersc.gov
[4] http://www.nersc.gov
[5] http://www.nccs.gov/computing-resources/jaguar/
[6] http://www.olcf.ornl.gov/titan/
[7] http://www.bw-grid.de

Computer	Peak Rank Top500	System	Perf. [PFlop/s]	Memory [TB]	Cores	Power [MW]	Usage [CPU hours]
bwGrid	-	IBM	-	22	2,828	-	> **900,000**
Kraken	3rd, Nov. 2009	Cray XT5-HE	1.03	147	98,928	3.09	> **90,000**
Franklin	8th, Nov. 2008	Cray XT4	0.36	75.5	38,642	1.15	> **1,000,000**
Hopper	5th, Nov. 2010	Cray XE6	1.28	217	153,408	2.91	> **5,000,000**
Jaguar	1st, Nov. 2009 and June 2010	Cray XK6	1.94	584	298,592	5.14	> **10,000**
Titan	1st, Nov. 2012	Cray XK7	17.59	693.5	560,640	8.2	> **30,000**

Table 4.1: Overview of the supercomputer sites, which were used for this and follow-up research.

The whole is greater than the sum of its parts.

Aristotle 384 BC-322 BC

Chapter 5

Role of Cel7A CBM

The research presented in this Chapter is based on the paper entitled "Simulation Analysis of the Cellulase Cel7A Carbohydrate Binding Module on the Surface of the Cellulose Iβ." and was submitted to the journal "Cellulose".

The cellulase enzyme Cel7A is a multi-domain enzyme consisting of a carbohydrate-binding module (CBM) and a catalytic domain (CD), joined by a linker peptide (Chapter 2). Recent experiments have also suggested that the activity of Cel7A may be limited by the accessibility to the cellulose substrate [183, 184], which is likely mediated by the CBM. Several studies aimed to understand the interaction of the CBM with the cellulose fiber [40–42,55,185–188]. Work using electron microscopy [40], single molecule fluorescence [41], atomic force microscopy [42], and computational docking [186] has shown that the Cel7A CBM binds preferentially to the hydrophobic faces of cellulose Iα. However, the nature of the CBM-cellulose binding is still not well understood. For example, while in Ref. [189] it was concluded that the binding of the CBM of the exoglucanase Cex from *Cellulomonas fimi*, which is similar to the CBM used in this thesis, to insoluble crystalline cellulose is entropically driven, in Ref. [190], on the other hand, it was suggested that the interaction of the CBM with non-crystalline cellulose is mainly enthalpically driven. To address this questions I performed simulations to understand the molecular mechanisms by which the CBM recognizes and interacts with specific fiber surfaces of cellulose Iβ. The simulation system chosen comprises the CBM of Cel7A and a 36-chain cellulose Iβ fiber. Reasons for studying only the CBM, and not the entire Cel7A, are to enable full convergence of the simulations and because the initial binding of Cel7A to the substrate is mainly mediated by the CBM [58–60]. Also, it has been shown experimentally that an isolated CBM can bind to cellulose crystals [40–42].

To overcome this time scale problem (Section 4.1), I have adopted a two-step simulation strategy. In the first step, an extensive set of coarse-grained all-atom BD simulations (>7600 trajectories, each 4 s long) are first used to probe the CBM-fiber diffusional encounter process from relatively large separation distances. The CBM was initially placed in random orientations and positions around the cellulose fiber. An aggregate of >29 ms of BD data was used to identify fiber faces (sites) and

CBM binding orientations for the most favorable CBM-fiber interactions. Based on the BD trajectories, a MSM was built to capture both thermodynamic and kinetic properties of the CBM binding sites on different fiber faces. In a second step, to study the detailed CBM-fiber interactions, the site-specific binding results from the BD simulations are then refined via multiple all-atom MD simulations to characterize the detailed CBM-fiber interactions. The MD simulations are started from the most favorable CBM binding configurations identified from the BD simulations, and have a combined simulation time of >1 μs. Using both simulation methods, my combined results provide a molecular-level description of the full CBM-cellulose fiber binding process, with important resulting implications for understanding the hydrolysis of crystalline cellulose by cellulases. An implicit treatment of the solvent, together with the rigid-body representation of the solutes, significantly reduces the computational cost, thus enabling BD simulation to be extended to much longer time and length scales than those accessible to MD.

In summary, the BD simulations probe the fiber faces to which the CBM binds, the binding orientation, and the kinetic properties of the observed binding states. In the MD simulations, properties influenced by explicit solvent or the internal flexibility of the systems are investigated, such as the detailed interactions of the CBM with the cellulose fiber, permitting the examination of conformational aspects of the CBM (e.g., the side chains of conserved Tyr residues) and the cellulose fiber (e.g., hydroxymethyl groups), hydrogen-bond patterns and hydration at the CBM-fiber interfaces, and the local diffusion of the CBM on the fiber surfaces.

5.1 Simulation and Analysis Details

5.1.1 BD and MD Simulations

As listed in Table 5.1, fourteen sets of BD simulations, referred to as BD1 to BD14, were performed. While BD1 and BD2 contain only the CBM or the cellulose fiber, respectively, BD3 to BD14 contain both the CBM and the cellulose fiber. In computer simulations, individual interaction terms can be selectively turned on/off, thus providing means of probing the roles of individual energy terms in influencing the thermodynamics and kinetics of the CBM-fiber binding. No intermolecular interaction was present in BD1 to BD3, BD8, and BD13, so these can be considered as free diffusion simulations. BD5 to BD12, and BD14 varied in the use of not only different interaction terms, but also different initial configurations. To cover the relevant CBM-fiber interaction space, the CBM was rotated randomly and its center placed in a random position in one of three shells around the fiber (Figure 4.5). 300 BD simulations were started from the first r_R shell (BD5 to BD8, and BD14), 500 from the second r_O shell (BD5 to BD8, and BD14) and 1000 from the third rs shell (BD9 to BD12) (Figure 4.5, bottom).

As listed in Table 5.2, twelve sets of MD simulations were performed. While MD1 and MD2 contain only the CBM or the cellulose fiber, respectively, MD3 to MD12 contain both the CBM and the cellulose fiber. MD3 to MD12 were all started from the most favorable CBM-fiber binding orientations obtained from the BD simulations, with the CBM placed in the middle of the fiber long axis and 3.5 Å above the fiber surfaces. This CBM-fiber distance of 3.5 Å was the most probable binding distance between the CBM bottom surface and the hydrophobic surfaces of the cellulose fiber from the BD simulations (consistent with the $C_\alpha - C_1$ distance probability profile shown in Figure 5.9), and was also used in a previous MD study [43]. For consistent comparison, the same initial distance was chosen for the MD simulations of the CBM on all fiber faces.

5.1.2 Orientation of Tyr Ring

The orientation of a tyrosine phenol ring relative to the cellulose fiber surface can be described by the Cartesian components x, y, and z of the two normalized vectors $\mathbf{p_{Tyr,1}}$, pointing from C_γ to C_ζ, and $\mathbf{p_{Tyr,2}}$, pointing from $C_{\epsilon 1}$ to $C_{\epsilon 2}$ (Figure 5.7b). In a spherical polar coordinate system, each of the two normalized vectors is defined by an azimuth ($-\pi < \alpha \leq \pi$) and an inclination angle ($0 \leq \beta \leq \pi$) (Figure 5.7c).

$$\alpha = \begin{cases} arctan\ (x/z) & \text{for } z > 0 \\ sgn\,(x) \cdot \pi/2 & \text{for } z = 0 \\ arctan\,(x/z) + \pi & \text{for } z < 0 \text{ and } x \geq 0 \\ arctan\,(x/z) - \pi & \text{for } z < 0 \text{ and } x < 0\,, \end{cases} \tag{5.1}$$

$$\beta = arccos\,(y)\ . \tag{5.2}$$

The orientation distribution of the Tyr ring is then visualized via an $\alpha - \beta$ map, the total volume under which is normalized to 1. The $\alpha - \beta$ map can be subdivided into different regions, each corresponding to a specific orientation of the Tyr ring towards the cellulose fiber surface (Figure 5.7a). $\alpha = 0°$ or $180°$ and $\beta = 90°$ denote a orientation parallel to the cellulose fiber axis, while $\alpha = \pm 90°$ and $\beta = 90°$ correspond to a orientation perpendicular to the fiber axis.

5.1.3 Hydrogen Bond Analysis

A hydrogen bond $X - H \dots Y$ is deemed to exist if the distance between atoms X and $Y \leq 3.5\,\text{Å}$ and the angle between the three atoms $X - H - Y \leq 30°$. The hydrogen bond occupancy for a given CBM atom in trajectory j is defined as the sum of the simulation times t_i when the hydrogen bond exists divided by the total simulation time $t_{sim,j}$

$$\tilde{o}_H(j) = \sum_k \frac{\sum_i t_i}{t_{sim,j}} \,, \tag{5.3}$$

where k sums over all the hydrogen bonds that a given CBM atom forms with different cellulose atoms. The hydrogen bond lifetime $\widetilde{t}_H(j)$ for a given CBM atom in trajectory j is defined as

$$\tilde{t}_H(j) = \frac{1}{n_j} \sum_i t_i \,, \tag{5.4}$$

n_j counts how many times the CBM-fiber hydrogen bond forms and breaks. The average occupancy o_H and lifetime t_H of a hydrogen bond over all the trajectories are given as the weighted means

$$o_H = \frac{\sum_j \lambda_j \cdot \tilde{o}_H(j)}{\sum_j \lambda_j} \,, \tag{5.5}$$

$$t_H = \frac{\sum_j \lambda_j \cdot \tilde{t}_H(j)}{\sum_j \lambda_j} \,, \tag{5.6}$$

where the weight factor λ_j is the length of each simulation trajectory j. λ_j is 12 for MD3 to MD6 and 40 for MD7 to MD12. The uncertainties of o_H and t_H are given by the weighted standard deviations of both quantities obtained from different simulation sets.

Name	Biomolecule	Number of trajectories	Length of single trajectory (total simulation time)	Interaction terms	Starting shell
BD1	CBM	1	10 μs	free diffusion	-
BD2	36-chain	1	10 μs	free diffusion	-
BD3	CBM and 36-chain	1	10 μs	free diffusion	-
BD4	CBM and 36-chain	1	55 μs	ES, VDW	ring
BD5	CBM and 36-chain	800	4 μs (3.2 ms)	ES, VDW	ring, octagon
BD6	CBM and 36-chain	800	4 μs (3.2 ms)	ES	ring, octagon
BD7	CBM and 36-chain	800	4 μs (3.2 ms)	VDW	ring, octagon
BD8	CBM and 36-chain	800	4 μs (3.2 ms)	free diffusion	ring, octagon
BD9	CBM and 36-chain	1000	4 μs (4 ms)	ES, VDW	ring segment (1, 0, 0)
BD10	CBM and 36-chain	1000	4 μs (4 ms)	ES	ring segment (1, 0, 0)
BD11	CBM and 36-chain	1000	4 μs (4 ms)	VDW	ring segment (1, 0, 0)
BD12	CBM and 36-chain	1000	4 μs (4 ms)	ES, VDW	ring segment (1, 0, 0)
BD12	CBM and 36-chain	1000	4 μs (4 ms)	ES, VDW	ring segment (0, 1, 0)
BD13	CBM and 36-chain	1	4 μs (3.2 ms)	free diffusion	-
BD14	CBM and 36-chain	800	4 μs (3.2 ms)	ES, VDW, HD, ED	ring, octagon

Table 5.1: Description of the different BD simulation sets. The simulations were started from three different shells (Figure 4.5). The interaction between the CBM and the 36-chain cellulose fiber is modeled as a sum of electrostatic (ES), van der Waals interaction (VDW), electrostatic desolvation (ED), and hydrophobic desolvation (HD) terms.

Name	Biomolecule	CBM docking face	Number of trajectories	Length of single trajectory (total simulation time)	Simulation box size	Number of atoms
MD1	CBM	-	1	120 ns	72 × 57 × 65 Å	24918
MD2	36-chain	-	1	120 ns	240 × 74 × 68 Å	115537
MD3	CBM and 36-chain	(1, 0, 0) hydrophobic	10	12 ns (120 ns)	240 × 100 × 90 Å	209587
MD4	CBM and 36-chain	(-1, 0, 0) hydrophobic	10	12 ns (120 ns)	240 × 100 × 90 Å	209587
MD5	CBM and 36-chain	(0, 1, 0) hydrophilic	10	12 ns (120 ns)	240 × 100 × 90 Å	209587
MD6	CBM and 36-chain	(0, -1, 0) hydrophilic	10	12 ns (120 ns)	240 × 100 × 90 Å	209587
MD7	CBM and 36-chain	(1, 0, 0) hydrophobic	2	40 ns (80 ns)	240 × 100 × 90 Å	209587
MD8	CBM and 36-chain	(-1, 0, 0) hydrophobic	2	40 ns (80 ns)	240 × 100 × 90 Å	209587
MD9	CBM and 36-chain	(0, 1, 0) hydrophilic	2	40 ns (80 ns)	240 × 100 × 90 Å	209587
MD10	CBM and 36-chain	(0, -1, 0) hydrophilic	2	40 ns (80 ns)	240 × 100 × 90 Å	209587
MD11	CBM and 36-chain	(-1, 1, 0) mixed	2	40 ns (80 ns)	240 × 100 × 90 Å	209587
MD12	CBM and 36-chain	(1, -1, 0) mixed	2	40 ns (80 ns)	240 × 100 × 90 Å	209587

Table 5.2: Description of the different MD simulation sets of the CBM and the 36-chain cellulose fiber.

5.2 BD Simulations of CBM and Cellulose Fiber Encounter Process

5.2.1 Identification of CBM Binding Faces

To chart how the CBM broadly encounters the cellulose fiber, I computed the density maps of the CBM around the fiber for the simulation sets BD5 to BD8, and BD14. These sets each covered all fiber faces and used identical simulation settings apart from variation of the potential energy terms (Figure 4.5 and Table 5.1). It is evident from the density map for BD5 (Figure 5.1a), which used the full potential, that high-density regions are present near the two hydrophobic faces and one mixed face of cellulose Iβ, indicating that the binding of the CBM favors the hydrophobic over the hydrophilic fiber faces: the CBM binds three to seven times more often to the two hydrophobic than to the two hydrophilic faces, consistent with previous experimental findings on cellulose Iα.

As a control, I also computed the density map for the free diffusion simulation BD8, which results, as expected, in a roughly uniform distribution over all fiber faces (Figure 5.1d). In the limit of infinite sampling, the free diffusion simulation should result in equal probability of CBM occupancy on all fiber positions, and thus the sampling error in the BD simulations can be estimated to be $\approx 1.5\,k_BT$. In the other BD simulations presented, the interaction forces between the CBM and the cellulose fiber were described by VDW and/or ES terms between individual atoms. The ES interaction also included a solvent polarization effect described by an effective charge model based on the classical Poisson-Boltzmann [191]. To assess the role of these individual energy terms in the CBM-fiber encounter process, I computed the density maps for BD6 and BD7 in which either the VDW or the ES interactions were turned off, respectively. As shown in Figure 5.1b and c, both these corresponding density maps show no preferential binding for any fiber face, indicating that both the VDW and the ES interactions contribute to the observed preferential binding of the CBM to the hydrophobic faces of the cellulose fiber.

To further understand the driving force underlying the observed preferential CBM-hydrophobic face binding, I calculated the VDW and ES energy components during the CBM-cellulose encounter process for the following four cases, in which the CBM is (1) close to ($\leq 15\,\text{Å}$) the hydrophobic fiber face, (2) far from ($\leq 40\,\text{Å}$) the hydrophobic fiber face, (3) close to ($\leq 15\,\text{Å}$) the hydrophilic fiber face and (4) far from ($\leq 40\,\text{Å}$) the hydrophilic fiber face. As shown in Figure 5.2, both the VDW and ES interaction energies become more favorable when the CBM moves closer to either fiber face.

The ES distributions do not show a notable difference between the hydrophilic and hydrophobic fiber faces, regardless of the CBM being close to or far away from the cellulose fiber. The VDW interaction is short-ranged in nature. Therefore, at far distances the VDW distributions for the hydrophilic and hydrophobic fiber faces (dashed lines) overlap with each other and both are close to zero. However, when the CBM comes closer to the fiber faces, the VDW interaction of CBM with the hydrophobic fiber face (blue solid line in Figure 5.2) is $\approx 8\,\text{kcal/mol}$ more favored than that with the hydrophilic fiber face (red solid line in Figure 5.2). These results suggest that both ES and VDW interactions contribute to the diffusional encounter of CBM with the cellulose fiber, but that the VDW interactions dictate the observed preferential binding of the CBM to the hydrophobic fiber faces.

The converged density map describes the binding probability of the CBM to the cellulose fiber, which is related to an equilibrium binding constant, K. K can be characterized by the ratio between an on-rate constant k_{on} and an off-rate constant k_{off}, i.e., $K = k_{on}/k_{off}$. I consider below how the individual VDW and ES interactions influence k_{on}, k_{off}, and thus K as a whole, on different fiber faces. I define outer and inner surface boundaries, at $(\leq 40\,\text{Å})$ and $(\leq 15\,\text{Å})$, respectively, from the fiber surface C_1 atoms (Figure 5.9). These two surface boundaries were chosen such that a sufficient number of CBM trajectories would cross them. I calculated the fraction of the BD trajectories in which the CBM (the C_α atoms of Y5, Y31, and Y32) started from the outer surface boundary and diffused to reach and bind to the cellulose surface as a function of time. k_{on} was then estimated by using the first-passage time, τ_{on} - the earliest time when the CBM crossed the inner surface boundary. Likewise, for the calculation of k_{off}, only those trajectories in which the CBM started from the inner surface boundary were included. I first calculated the fraction of the trajectories that escaped from the inner surface boundary and reached the outer surface boundary as a function of time. k_{off} was then estimated by using the first-passage time, τ_{off}.

As listed in Table 5.3, k_{on} is not very different between the hydrophobic and the hydrophilic faces, while k_{off} increases from $(0.14 \pm 0.09) \cdot 10^{-6}\,\text{ps}^{-1}$ to $(0.58 \pm 0.19) \cdot 10^{-6}\,\text{ps}^{-1}$ when changing from the hydrophobic to the hydrophilic faces, leading to a reduction of the CBM binding constant K by a factor of 5 on the hydrophilic face relative to the hydrophobic face. This result is quite consistent with the above density map analysis. On the hydrophobic faces, k_{off} increases by two orders of magnitude when turning off either the ES or the VDW interaction. In comparison, k_{on} is only marginally affected by turning off the VDW interaction, but is reduced by a factor of 9 when turning off the ES interaction. Taken together, the preferential binding of CBM to the hydrophobic surfaces arises mainly from the different dissociation rates, k_{off}. For the binding of the CBM to the hydrophobic fiber face, k_{off} is influenced by both the VDW and ES interactions, while k_{on} is more affected by the ES interaction than the VDW interaction. The ES interaction accelerates the CBM binding from far distances, while the VDW interaction becomes more dominant at short distances and keeps the CBM more tightly bound to the hydrophobic faces (Figure 5.2).

5.2.2 Orientations of CBM on Hydrophobic Fiber Faces

To understand in more detail the CBM-cellulose fiber interaction, I analyzed the orientations of the CBM relative to the fiber surfaces for those CBM configurations located within the inner surface boundary of the hydrophobic faces. The analysis shows that the CBM preferentially adopts a binding pose in which the hydrophobic patch of the CBM stacks against the fiber face, with a probability of $57 \pm 1\%$. The stacked configurations were defined as those with the CBM hydrophobic patch (i.e., all the heavy atoms in the phenol rings of Y5, Y31, and Y32) located within 8 Å of the fiber surface C_1 atoms. The stacked configurations were further analyzed by examining the alignment of the CBM with respect to the long fiber axis, which was described by an angle α_{BD} between the fiber axis and a line passing through the three phenol ring centers of Y5, Y31, and Y32 (Figure 2.13). The probability distribution of α_{BD} for all the stacked CBM configurations is displayed in Figure 5.3, and shows that the parallel (0°) and the anti-parallel (180°) orientations dominate. Both parallel and anti-parallel orientations increase the CBM-fiber contacting surface by $\approx 52\,\%$ compared to the perpendicular orientation, providing an explanation for this preference.

Integration of the probability distribution of α_{BD} in Figure 5.3, from 0°to 90°for the parallel binding and from 90° to 180° for the anti-parallel binding, reveals that the anti-parallel binding is slightly favored over the parallel one, with probabilities of $56 \pm 1\,\%$ vs. $44 \pm 1\,\%$. Both the CBM and the fiber have non-zero dipole moments: that of the CBM is 72 D, aligned roughly along the line connecting the three C_α atoms of Y5, Y31, and Y32, while the cellulose fiber has a dipole moment of 489 D oriented along the fiber axis. Therefore, it is reasonable to speculate that the dipole-dipole interaction may account for the observed preference for the anti-parallel binding.

If the dipole-dipole interaction plays a role in the preferential anti-parallel binding, a gradual anti-parallel alignment of the CBM relative to the fiber might be expected during the encounter process. To test this idea, I calculated the probability distributions of the α_{BD} angle during the BD simulations, and from that the fraction of the parallel and anti-parallel orientations. As shown in Figure 5.3, the fraction of the anti-parallel distribution $A_{AP}(t)$ indeed increases slowly over the time, reaching a plateau at $56 \pm 1\,\%$ after about 70 ns. Completely anti-parallel alignment of the CBM was not observed in the BD simulations due in part to the CBM either escaping from or coming too close to the fiber. In the latter case, the dipolar interaction approximation broke down and local non-bonded (especially VDW) interactions started to dominate, and these do not distinguish between the parallel and the anti-parallel modes of binding.

Most exocellulase enzymes show remarkable directional specificity for their processive activity. For instance, Cel7A hydrolyzes cellulose from the reducing end, while Cel6A proceeds from the non-reducing end. This directional specificity has been thought to arise from the structural arrangement of the enzyme active site allowing the chemical reaction to proceed only along a specific direction [49, 192–194]. The present results suggest that the preferential anti-parallel CBM-fiber binding arising from the dipole-dipole interaction may also potentially contribute to the directional specificity of cellulase enzymes. The interplay between the CD and the CBM via the linker at the molecular level is not fully characterized [62, 195–198], and in particular it is unclear whether the CD pushes the CBM [77, 197] or the CD pulls the CBM during the processive hydrolysis of crystalline celluloses by harnessing the chemical energy released from the cleavage of the glycosidic bond. The anti-parallel CBM-fiber binding would support the second mechanism.

5.2.3 Characterization of CBM-Fiber Docking States

In order to provide a more quantitative description of the observed binding states of the CBM on the fiber surfaces and their kinetic relationship, I built a MSM for the CBM-fiber encounter process. For this, the cross-sectional area (the yz plane) around the fiber was discretized into 100 states, which were then combined and mapped onto individual fiber faces (Section 2.1 and Section 4.6). Figure 5.4 shows the stationary distribution probability, π_i, which gives the occupancy probability of the CBM at a given fiber position i, and the mean exit time, $t_{exit,i}$ which quantifies the length of time the CBM stays bound to fiber position i before escaping. The combined distribution probability and mean exit time together with the transition probability, t_{ii}, for each fiber face are displayed in Figure 5.4. The hydrophobic fiber faces (1, 0, 0) and (-1, 0, 0) show averaged stationary distribution probabilities of $0.57 \pm 0.2\,\%$ and $0.41 \pm 0.1\,\%$, respectively, as compared with that of $0.1 \pm 0.1\,\%$ for both hydrophilic faces (0, 1, 0) and (0, -1, 0), confirming that the hydrophobic faces are more likely to be bound by the CBM than the hydrophilic faces. Moreover, once reaching a given hydrophobic face, the CBM is more likely to stay bound, with transition probabilities of $12 \pm 1\,\%$ and $16 \pm 1\,\%$ compared to the hydrophilic faces which have probabilities of $10 \pm 2\,\%$ and $9 \pm 2\,\%$. The corresponding mean exit times are 1.4 ± 0.1 ns and 1.6 ± 0.1 ns for the hydrophobic faces, compared to 1.3 ± 0.1 ns and 1.2 ± 0.1 ns for the hydrophilic faces. These MSM results broadly agree with those from the above density map and kinetic analyses.

5.3 MD Simulations of Interaction of CBM with Fiber Surfaces

Explicit water molecules and the internal flexibility of the CBM and the fiber are not modeled in the BD simulation. This limitation of the BD approach is overcome in the all-atom MD simulation, which provides a way to investigate the ultimate details of how individual atoms move and which motions may be linked to biological functions.

5.3.1 Contacts between CBM and Fiber Surfaces

To quantitatively examine the interaction of the CBM with the fiber surfaces, the distance distributions $p(d)$ between the heavy atoms in the phenol rings of Y5, Y31, and Y32 and those on the fiber surfaces were calculated, which are plotted in Figure 5.5. The mean distances between the CBM hydrophobic patch (consisting of Y5, Y31, and Y32) and the hydrophobic faces are shorter than for the hydrophilic fiber faces. Moreover, all the $p(d)$ profiles for the hydrophobic fiber faces show a single, prominent peak, while those for the hydrophilic fiber faces are much more broadly distributed. Furthermore, the distance distributions for different carbon atoms in the same Tyr ring peak at slightly shifted positions, indicating the three Tyr rings being tilted relative to the fiber surface, which will be further discussed below. The observed closer binding distances of Y5, Y31, and Y32 to the hydrophobic fiber faces, together with their sharper distributions on these surfaces, indicate that the CBM interacts more strongly with the hydrophobic fiber faces.

5.3.2 Hydroxymethyl Group Conformations

The hydroxymethyl groups on the fiber surfaces are directly involved in the interaction of the cellulose fiber with the solvent and cellulolytic enzymes. It has been demonstrated that the formation of the two intramolecular hydrogen bonds, $O_3H - O_5$ and $O_2H - O_6$, critically depends on the particular conformation of the hydroxymethyl group [38]. The hydroxymethyl conformation is defined by the dihedral angle χ between O_5, C_5, C_6, and O_6 atoms of a pyranose residue. χ clusters around 180°, 60°, and -60°, with the corresponding conformations named *trans*, *gauche*$^+$, and *gauche*$^-$, respectively. Figure 5.6 shows a comparison between the χ probability distributions for the different cellulose fiber surfaces that are in contact with the CBM and those for a cellulose fiber only system (simulation MD2). Consistent with the cellulose Iβ structure, all hydroxymethyl groups started from a trans conformation. As the hydrophilic fiber faces interact relatively weakly with the CBM, the hydroxymethyl groups were free to rotate to the more stable *gauche*$^+$ conformation, as also observed in the MD simulation containing only the cellulose fiber (MD2). In contrast, as the hydrophobic fiber faces interact more strongly with the CBM, the hydroxymethyl groups were rotationally hindered and did not rotate as freely away from the initial trans conformation.

5.3.3 Tyrosine Side Chain Conformations

To characterize in further detail the interactions of the CBM with the fiber faces of different hydrophobicity, I examined the binding conformations of the Tyr side chains on the cellulose fiber surfaces by computing the two torsional angles $\chi_{Tyr,1} = (\mathrm{C}, \mathrm{C}_\alpha, \mathrm{C}_\beta, \mathrm{C}_\gamma)$ and $\chi_{Tyr,2} = (\mathrm{C}_\alpha, \mathrm{C}_\beta, \mathrm{C}_\gamma, \mathrm{C}_{\delta 2})$. In the simulation of an isolated CBM (MD1), the probability distributions of both $\chi_{Tyr,1}$ and $\chi_{Tyr,2}$ for Y5, Y31, and Y32 are all split into two peaks, except for $\chi_{Tyr,1}$ of Y32 that exhibits a single sharp peak (Figure 5.6). When the CBM is bound to the hydrophobic fiber surfaces, nearly all the Tyr torsional angle distributions show only one single peak, except for $\chi_{Tyr,2}$ of Y5 that shows a small second peak. In contrast, on the hydrophilic fiber surfaces, the $\chi_{Tyr,2}$ profiles of Y5, Y31, and Y32 all show two distinct peaks, very similar to what is observed in the simulation of an isolated CBM. These results also show that the probability distributions are generally more sharply peaked on the hydrophobic faces than on the hydrophilic faces, suggesting the rotation of the three tyrosine side chains is more restricted on the hydrophobic faces. These results are in agreement with the distribution profiles of $\chi_{Tyr,1}$ and $\chi_{Tyr,2}$ for Y5 reported previously [186], although the present study has systematically shown the distribution profiles of the three tyrosine side chains on all fiber faces.

The orientations of the three Tyr rings (Y5, Y31, and Y32) with respect to the fiber surfaces are characterized by the $\alpha - \beta$ maps of the two vectors $\chi_{Tyr,1}$ and $\chi_{Tyr,2}$, and these are displayed in Figure 5.7. Again, the $\alpha - \beta$ maps are more sharply peaked for the hydrophobic than for the hydrophilic faces, indicating that the stack orientations (β is centered at $\approx 90°$) on the hydrophobic surfaces are more discriminating. The tilting of a Tyr phenol ring towards the fiber face can be described by the inclination angle β of the vector $\chi_{Tyr,1}$ or $\chi_{Tyr,2}$ (Figure 5.7c), with the tilt angle δ defined by $|\beta - 90°|$. As shown in Figure 5.7d, on the hydrophobic faces the three Tyr rings are all slightly tilted, by 6.0° to 10.3° (Table 5.9), such that the O_ξ atom is closer to the fiber surfaces than the $C\gamma$ atom. In particular, the tilting of Y31, which borders the CBM bottom and wedge faces, substantiates its possible role in helping load a detached cellulose chain over the CBM wedge face [55]. The alignment of a Tyr phenol ring relative to the fiber long axis can be described by the azimuth angle α of the vector $\mathbf{p}_{Tyr,1}$ or $\mathbf{p}_{Tyr,2}$. As shown in Figure 5.7d, on the hydrophobic faces, the Y31 ring orients parallelly (pointing to the non-reducing end) while Y32 anti-parallelly (pointing to the reducing end) relative to the cellulose chain. Y5 is located at the rear end of the CBM and therefore its orientation is less restricted, consistent with the side chain torsional angle analysis provided in Figure 5.6. In the course of the simulations (MD7 and MD8), the three Tyr rings orient in such a way that the contact surface area is maximized to enhance the interaction between the CBM and the hydrophobic fiber surfaces.

5.3.4 Hydration at CBM-Fiber Interfaces

To investigate possible roles of dewetting in the CBM-fiber binding, I examined the hydration between the CBM and the fiber surfaces by counting the number of water molecules within 4.5 Å of the ring carbon atoms of Y5, Y31, and Y32. A significantly smaller number of solvent molecules were found for the hydrophobic than for the hydrophilic fiber faces (Table 5.3). However, a complete dewetting between the CBM and the hydrophobic fiber surfaces did not occur, possibly because the CBM hydrophobic patch is relatively small and surrounded by many polar residues, such as H4, Y5, Q7, N29, Y31, Y32, and Q34. The observed partial dewetting phenomenon resembles what has been observed previously for peptide folding and protein-protein interactions in which where water molecules were removed from between the two hydrophobic surfaces [199–203]. The partial dewetting brings the CBM and the fiber surfaces closer, which in turn increases the probability of hydrogen bonding between the two macromolecules.

5.3.5 CBM-Fiber Hydrogen Bonds

While mutagenesis studies suggested that N29, Y31, and Q34 form hydrogen bonds with the cellulose fiber [55, 188, 204], MD simulations have found that Y5, Q7, N29, Y31, and Y32 form hydrogen bonds with the primary alcohol groups on the cellulose surface [62, 205]. Moreover, the MD study has shown that Q2, S3, and H4 also form hydrogen bonds with broken cellulose chain ends.

Here, to ensure a statistically meaningful comparison, only those hydrogen bonds with an occupancy $\geq 5\,\%$ are considered. According to this criterion, H4, Y5, Q7, I11, L28, N29, Y31, Y32, and Q34 are engaged in hydrogen bonding with the cellulose fiber, although at no time all these hydrogen bonds are present simultaneously (Table 5.4). Of these, N29 has the highest probability of intermolecular hydrogen bonding (Figure 5.8). H4, which has been shown to hydrogen bond with a broken chain [62], is found to also form hydrogen bonds with the intact cellulose fiber faces. It is interesting to note that N29 is more likely to form a hydrogen bond with the hydrophobic than with the hydrophilic fiber faces. Significantly higher hydrogen bond occupancy is also found for Q34 on the hydrophobic face (-1,0,0) than on the other faces. For all the other hydrogen bonds, no statistically significant difference is found between the hydrophobic and hydrophilic fiber faces. Overall, the CBM populates more hydrogen bonds with the hydrophobic than with the hydrophilic fiber faces (the average occupancy is $39.5 \pm 15.9\,\%$ larger). These results are consistent with previous computational [62, 205] and experimental studies [55, 188, 204].

5.3.6 CBM Diffusion on Fiber Surfaces

The motion of the CBM on the cellulose fiber surface can be approximately described as a Brownian diffusion [206–208], which can then be quantified by the translational diffusion coefficient $D_{trans}^{CBM,x}$. Assuming a Brownian diffusion along the x axis of the cellulose fiber, $D_{trans}^{CBM,x}$ was calculated from the mean square displacement of the

CMB using the Einstein relation in the x dimension,

$$D_{trans}^{CBM,x} = lim_{x \to \infty} \left\langle [x(t+t_0) - x(t_0)]^2 \right\rangle_{t_0} , \tag{5.7}$$

where $x(t)$ is the position of the center of the CBM at time t. The thus calculated diffusion coefficients of the CBM on the fiber surfaces are in the range of 0.25 to 2.63 Å^2/ns (Table 5.5), consistent with previous studies [209–211]. These values are of the same order of magnitude but smaller than that of the free CBM in solution obtained from MD1 (Table 5.6), indicating that the CBM diffusion is hindered when in close contact with the cellulose fiber. Also, I find that the diffusion constant is $\approx 90\,\%$ smaller on the hydrophobic faces than on the hydrophilic faces (Table 5.5).

5.4 Summary

The deconstruction of cellulosic biomass by Cel7A can be roughly subdivided into the steps (S1) to (S6) (Section 4.3). This thesis provides atomic-detail insights into steps (S1)-(S3), the BD concerning primarily (S1) and the MD (S2)-(S3), while steps (S4)-(S6) require QM simulations and are outside the scope of this thesis.

(S1) Initial Encounter of CBM with Cellulose Fiber. The BD results show a clear preference for the CBM to dock to the hydrophobic faces of the cellulose fiber, with both the electrostatic and van der Waals interactions being required for the observed preference. The most favorable docking orientation is a parallel stacking of the CBM bottom surface formed by Y5, Y31, and Y32 against the hydrophobic fiber faces, suggesting that the non-polar van der Waals interaction is a relevant driving force for the initial CBM-fiber encounter (Figure 5.2). At wide separation distances, the electrostatic interaction accelerates the diffusional encounter process, as evidenced by the decrease of k_{on} by one order of magnitude observed when turning off the electrostatic interactions in the BD simulations. Moreover, the dipole-dipole interaction between the CBM and the cellulose fiber also helps align the CBM antiparallel to the fiber axis, which may contribute to the directional specificity of the Cel7A enzyme [77,197] (Figure 5.3). Furthermore, the kinetic and MSM analyses indicate that the thermodynamic preference for the hydrophobic fiber faces of the CBM binding arises mainly from a slower k_{off} rate. That is, the hydrophobic fiber faces are stickier for CBM binding than are the hydrophilic ones.

(S2) Diffusion on Fiber Surface. Once docked to the cellulose fiber, the CBM diffuses on the fiber surfaces to locate the regions for better binding, i.e., the molecule makes a transition from a 3-dimensional search to a 2-dimensional local search. As listed in Table 5.5 and 5.6, the present MD results that the diffusion of the CBM on all the fiber surfaces is hindered, with the diffusion on the hydrophobic surfaces being more restricted than that on the hydrophilic ones. The calculated absolute effective velocities for the CBM of 0.36 ± 0.20 Å/ns on the hydrophilic surfaces and 0.10 ± 0.08 Å/ns on the hydrophobic surfaces broadly agree with those of 0.44 to 1 Å/ns obtained in a previous MD study [62], but are several orders of magnitude greater than the experimentally measured [212] sliding velocity of $3.5 \cdot 10^{-8}$ Å/ns of an intact Cel7A on crystalline cellulose. Possible reasons for this difference are two fold: (1) different structures were used in these studies, i.e., a heavier, intact Cel7A in the experiment vs. only an isolated CBM in the simulations and (2) the experimentally measured velocity corresponds to that of a reaction-coupled diffusion, in which other rate-limiting processes, such as chemical reaction, cellulose decrystallization and processive threading of a detached single cellulose chain inside the catalytic tunnel, may also be involved.

(S3) Binding of CBM to Fiber Surface. After having located a potential binding site on the hydrophobic fiber surfaces, the CBM moves closer to the fiber surfaces, which is coupled with partial removal of the interfacial water molecules. Meanwhile, the CBM adjusts its conformation to bind more strongly to the cellulose fiber, similar to a induced fit mechanism. The contact surface area between the CBM and the fiber is increased by stacking the Y31 and Y32 rings against the two cellulose pyranose rings, and extending them (towards the non-reducing and reducing ends for Y31 and Y32, respectively) along the fiber axis direction. The total number of hydrogen bonds between the CBM and the fiber surfaces is also increased to 14.1 ± 1.5. These hydrogen bonds are expected to play an important role in stabilizing the CBM-fiber complex and may also explain why the diffusion of the CBM is hindered on the fiber surfaces. The MD simulation results indicate that the hydrogen bonding occupancy is 39.5 ± 15.9 % greater for the hydrophobic than for the hydrophilic fiber faces, indicating that local polar interactions also contribute to the preferential binding of the CBM to the hydrophobic over the hydrophilic surfaces. Overall, the simulation results presented support the conclusion that the nature of the CBM-cellulose binding is neither purely hydrophobic nor purely polar, instead involving both types of interaction. The apparent contradiction in previous experiments [189, 190] can be reconciled by noting the fact that the hydrophobic interaction likely diminishes for the binding of the CBM to non-crystalline cellulose.

Experimental studies have shown that mutation of Y31 to His, Asp, or Ala reduces

the binding and activity of the Cel7A on crystalline cellulose to a level similar to an isolated CD [55]. The presented $\alpha - \beta$ maps suggest in the CBM-fiber complex the rotation of Y31 is less restricted than Y32, but more restricted than Y31 in the free CBM simulation (Figure 5.7). Moreover, the simulations suggest a conformational rearrangement of Y31, permitting it to stack against the hydrophobic fiber surfaces. It is also interesting to note that the phenol ring of Y31 is slightly tilted towards the fiber surface (Figure 5.7), and in this tilted position Y31 is more likely to hydrogen bond with the cellulose fiber (Figure 5.8). The lack of either stacking or hydrogen bonding interaction with the cellulose fiber might explain why the mutations of Y31 in the previous study [55] have reduced the binding activity of Cel7A on crystalline cellulose.
Y13 has been suggested to undergo a conformational change induced by a broken cellulose chain end, moving away from the protein interior to the wedge surface to make VDW contact with the broken cellulose chain [62]. In all the present simulations with an intact cellulose fiber (no broken chain), Y13 remains protruding out of the wedge surface, with an average solvent accessible surface area (SASA) [213] of > 69 Å^2 (Table 5.7). Y13 is even fully solvent exposed in 5 out of the 52 trajectories, with an SASA of 145.3 ± 13.8 Å^2, in comparison to an average SASA of 78.6 Å^2 to 152.4 Å^2 for the three surface-forming residues Y5, Y31, and Y32 (Table 5.7). Therefore, my simulations suggest that the observed conformational change of Y13 reported in Ref. [62] can also occur in the absence of a broken cellulose chain, although I cannot exclude the possibility that the solvent-exposed Y13 may be further stabilized by the broken chain end.

Overall, my simulation results provide atomic-level details of the encounter and binding of the CBM to the cellulose fiber, which are broadly consistent with a large body of experimental data on the CBM interaction with crystalline cellulose, and can serve as a basis for engineering improved CBMs for enzymatic deconstruction of insoluble celluloses.

BD simulations	k_{on} $[10^{-6}ps^{-1}]$	k_{off} $[10^{-6}ps^{-1}]$	k_{on}/k_{off}
ES+VDW (hydrophilic)	1.10±0.10	0.58±0.19	1.89
ES+VDW (hydrophobic)	1.30±0.30	0.14±0.09	9.20
ES (hydrophobic)	0.90±0.30	18.29±1.92	0.05
VDW (hydrophobic)	0.14±0.10	19.07±1.08	0.01

Table 5.3: Association and dissociation rates of the binding of the CBM to different cellulose fiber surfaces calculated from the CBM-fiber BD simulations.

	Number of water molecules			
CBM docking face	Y5	Y31	Y32	Y5, Y31, Y32
CMB only	3.4 ± 2.3	3.2 ± 2.3	2.4 ± 1.7	8.8 ± 4.0
(1, 0, 0) hydrophobic	2.0 ± 1.3	2.6 ± 1.4	0.6 ± 0.7	5.1 ± 2.1
(-1, 0, 0) hydrophobic	2.1 ± 1.3	2.2 ± 1.4	0.8 ± 0.9	5.1 ± 2.1
(0, 1, 0) hydrophilic	3.7 ± 1.8	2.7 ± 1.7	1.9 ± 1.3	8.2 ± 2.7
(0, -1, 0) hydrophilic	3.3 ± 1.9	2.3 ± 2.0	2.1 ± 1.7	7.7 ± 3.7
(-1, 1, 0) mixed	3.9 ± 1.9	2.1 ± 1.3	1.5 ± 1.3	7.1 ± 2.4
(1, -1, 0) mixed	3.5 ± 1.9	2.9 ± 1.9	0.7 ± 0.9	7.5 ± 2.6

Table 5.4: Number of water molecules within a radius of 4.5 Å around the residues Y5, 31, and 32 calculated for the MD simulations (MD1, and MD3 to MD12).

	Percentages of simultaneous hydrogen bonds (H4, Q7, N29, Q34)				
CBM docking face	0	1	2	3	4
(1, 0, 0) hydrophobic	3.1 %	30.2 %	45.1 %	19.6 %	2.0 %
(-1, 0, 0) hydrophobic	0.3 %	17.9 %	65.2 %	14.5 %	2.1 %
(0, 1, 0) hydrophilic	15.5 %	32.6 %	29.1 %	16.7 %	6.2 %
(0, -1, 0) hydrophilic	41.0 %	27.0 %	22.4 %	9.0 %	0.7 %
(-1, 1, 0) mixed	14.2 %	18.3 %	50.4 %	14.1 %	2.9 %
(1, -1, 0) mixed	3.1 %	21.7 %	45.7 %	27.3 %	2.2 %

Table 5.5: Overview of the probabilites of hydrogen bonds formed simultaneously between the cellulose fiber and the CBM residues H4, Q7, N29, and Q34 [61, 62].

CBM docking face	Set	$D_{trans}^{CBM,xyz}$ [Å^2/ns]	$D_{trans}^{CBM,x}$ [Å^2/ns]
(1, 0, 0) hydrophobic	MD3, MD7	0.21±0.18	0.25±0.14
(-1, 0, 0) hydrophobic	MD4, MD8	0.31±0.25	0.45±0.39
(0, 1, 0) hydrophilic	MD5, MD9	1.79±1.18	1.99±1.12
(0, -1, 0) hydrophilic	MD6, MD10	4.26±2.20	2.63±1.06
(-1, 1, 0) mixed	MD11	0.31±0.12	0.39±0.16
(1, -1, 0) mixed	MD12	0.80±0.21	0.48±0.26

Table 5.6: Translational diffusion coefficients of the CBM on different cellulose fiber faces. $D_{trans}^{CBM,xyz}$ denotes the diffusion of the CBM in three dimensions and $D_{trans}^{CBM,x}$ is the diffusion along the cellulose fiber x-axis.

	D_{trans} [$\AA^2/ns$]	D_{rot} [$10^{-3} rad^2/ns$]
free 36-chain	5.00	0.10
free Cel7A	5.50	2.39
free CD	7.01	5.14
free CBM	16.40	66.16

Table 5.7: Translational and rotational diffusion coefficients D_{trans} and D_{rot} for the free 36-chain cellulose fiber model, the entire Cel7A, the CD, and the CBM. The experimental estimated diffusion coefficient for the Family 2 CBM bound on the cellulose surface is $0.0002 \ldots 0.012$ [210].

		Average solvent accessible surface area [Å^2]			
CBM docking face	Set	Y5	Y13	Y31	Y32
CMB only	MD1	147.7 ± 13.8	72.0 ± 8.9	144.9 ± 13.5	83.0 ± 11.7
(1, 0, 0) hydrophobic	MD3, MD7	141.5 ± 13.6	69.0 ± 7.5	141.1 ± 10.0	78.6 ± 10.4
(-1, 0, 0) hydrophobic	MD4, MD8	147.4 ± 13.9	71.6 ± 8.8	142.4 ± 11.1	83.8 ± 11.2
(0, 1, 0) hydrophilic	MD5, MD9	141.8 ± 13.9	70.2 ± 8.3	143.4 ± 11.1	82.7 ± 10.6
(0, -1, 0) hydrophilic	MD6, MD10	142.1 ± 13.9	70.6 ± 8.3	143.9 ± 11.7	84.4 ± 12.0
(-1, 1, 0) mixed	MD11	146.6 ± 13.8	74.6 ± 10.5	145.1 ± 10.6	83.2 ± 10.3
(1, -1, 0) mixed	MD12	152.4 ± 13.8	70.8 ± 8.8	148.9 ± 13.4	84.4 ± 10.4

Table 5.8: Average solvent accessible surface area of the CBM residues Y5, Y13, Y31, and Y32.

CBM docking face	Set	Y5	Y31	Y32
(1, 0, 0) hydrophobic	MD3, MD7	10.3°±3.1°	6.0°±2.2°	6.3°±2.1°
(-1, 0, 0) hydrophobic	MD4, MD8	10.3°±3.1	10.0°±3.2°	6.3°±2.2°

Table 5.9: Tilt angle δ between the Tyr ring relative to the fiber surface plane ($\delta = |\beta - 90°|$ Figure 5.1c).

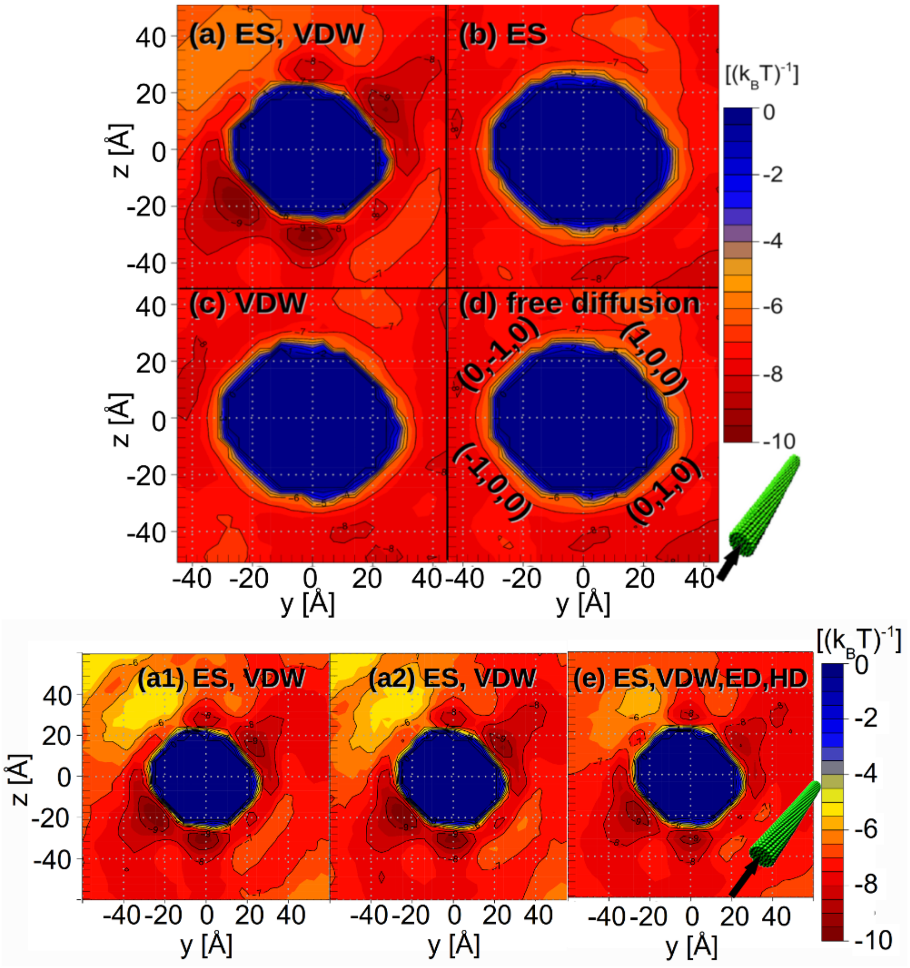


Figure 5.1: (Top) The CBM-fiber density maps for the BD sets BD5 to BD8 are shown in (a) to (d), respectively. (Bottom) The BD trajectories of the BD set BD5 were randomly split into two subsets, and the density map for each subset is shown in figure (a1) or (a2). The density maps for the BD set BD14 is shown in figure (e). The interaction terms include electrostatic (ES), van der Waals (VDW), hydrophobic desolvation (HD) and electrostatic desolvation (ED).

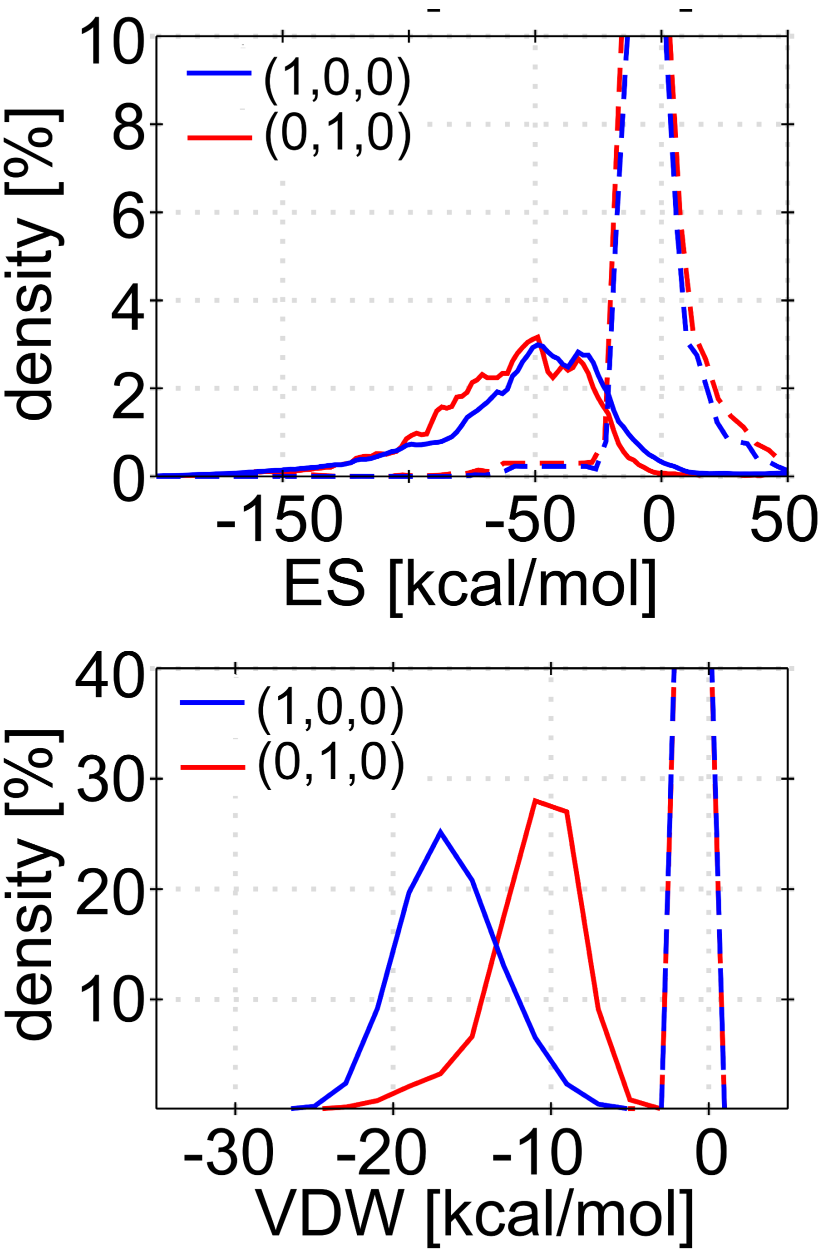


Figure 5.2: Distributions of ES (top) and VDW (bottom) CBM-fiber interaction energies for the hydrophobic (1, 0, 0) (blue lines) and the hydrophilic (0, 1, 0) (red lines) cellulose fiber faces. The areas under the distributions are normalized to 100 %. The distributions for the CBM close to the fiber face (< 15 Å) are plotted as solid lines and those for the CBM far from the fiber face (> 40 Å) as dashed lines. The cutoff distances are defined as between the C_1 atom of the fiber surface and the C_α atoms of Y5, Y31, and Y32 of the CBM.

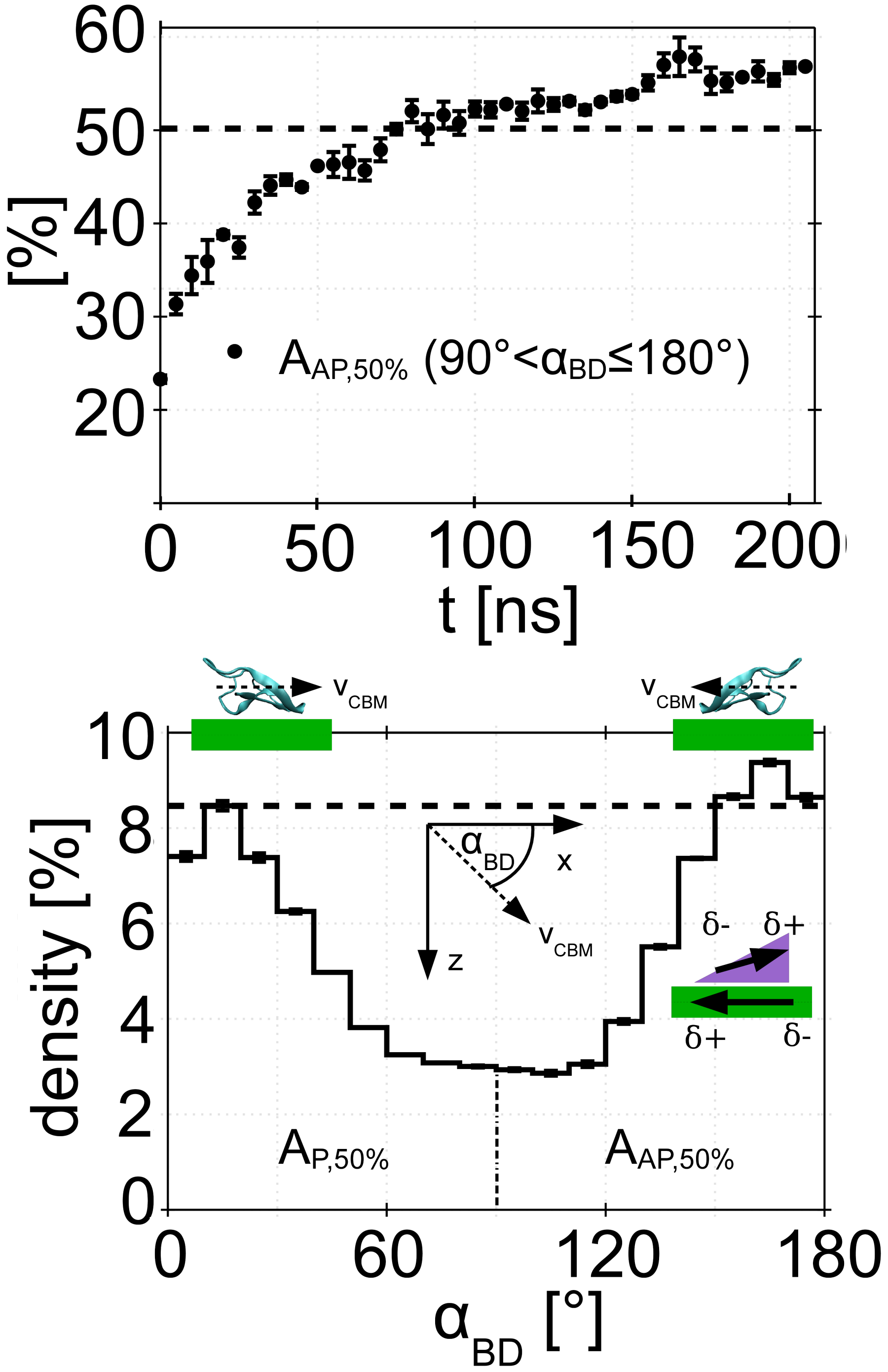


Figure 5.3: (Top) Probability of the anti-parallel alignment of the CBM with respect to the cellulose fiber as a function of time. (Bottom) Probability distribution of the orientational angle α_{BD} of the CBM relative to the fiber axis is shown for those CBM states with the Tyr rings stacked against the hydrophobic fiber faces. The area under the histogram is normalized to 100 %. A_P sums the histogram density from 0° to 90° (parallel alignment) and A_{AP} from 90° to 180° (anti-parallel alignment).

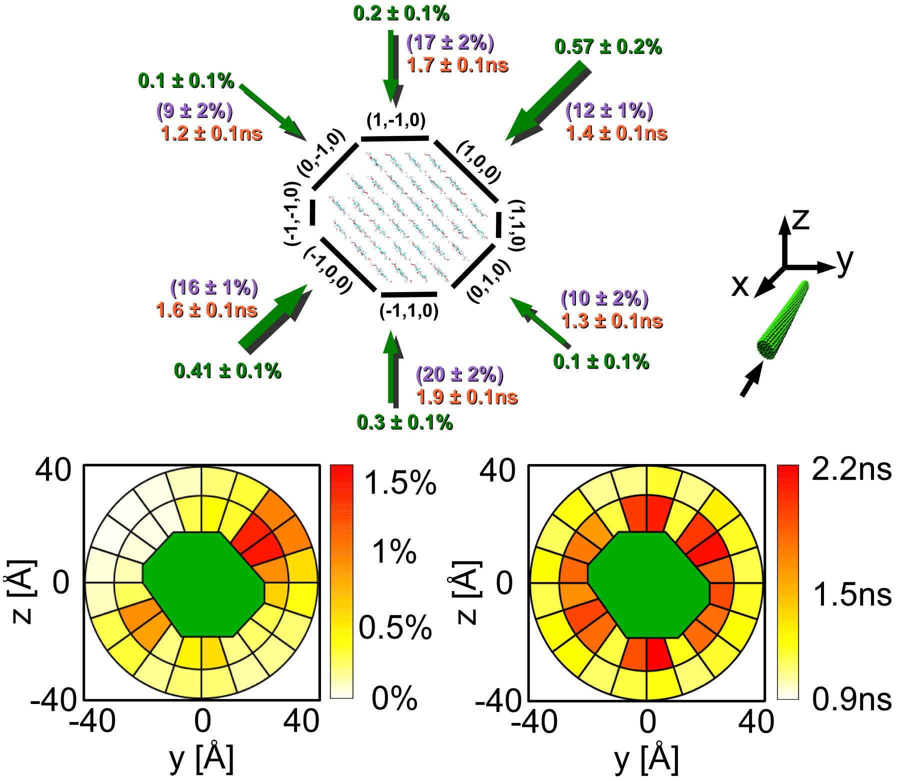


Figure 5.4: (Top) Averaged stationary distribution π_i (green), mean exit time $t_{exit,i}$ (red) and probability t_{ii} (violet) for all fiber faces. Bins at the corners cannot be associated clearly with any specific fiber face, thus are neglected in the analysis. Visualization of the stationary distribution π_i (bottom left) and mean exit time $t_{exit,i}$ (bottom right) for the CBM-fiber encounter process in BD dataset BD5.

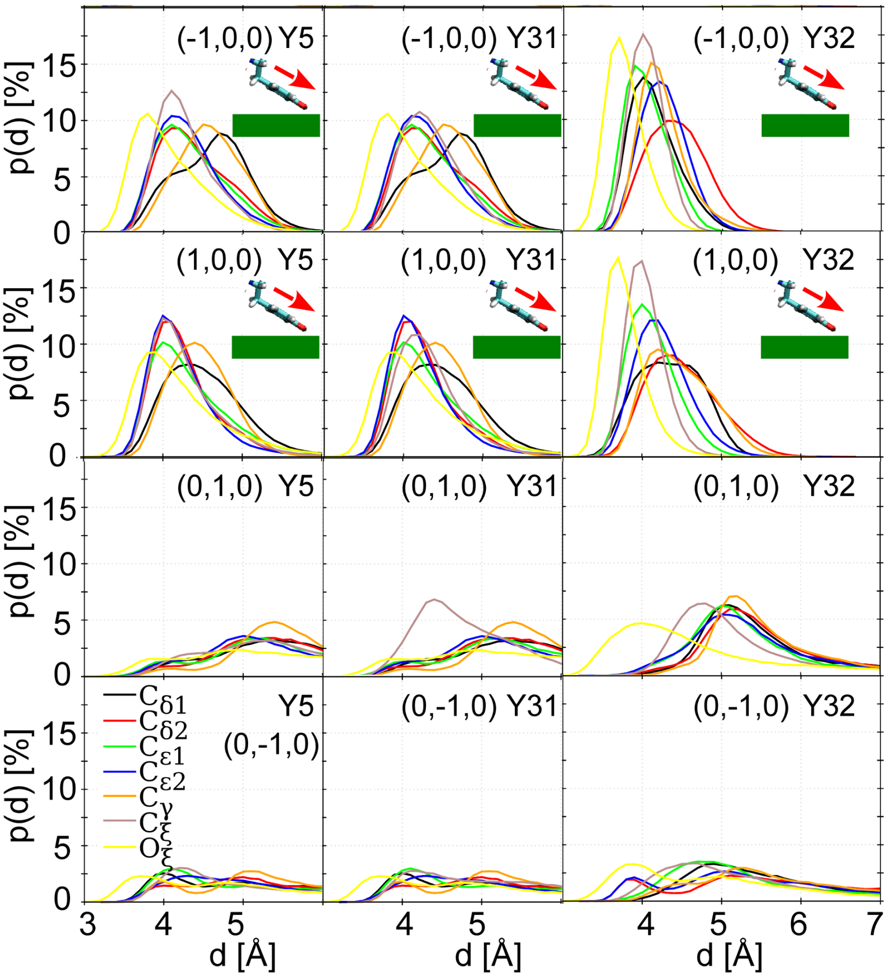


Figure 5.5: Probability distributions $p(d)$ of the distances between the non-hydrogen atoms of the fiber surfaces and those in the phenol rings of the three Tyr residues Y5, Y31, and Y32.

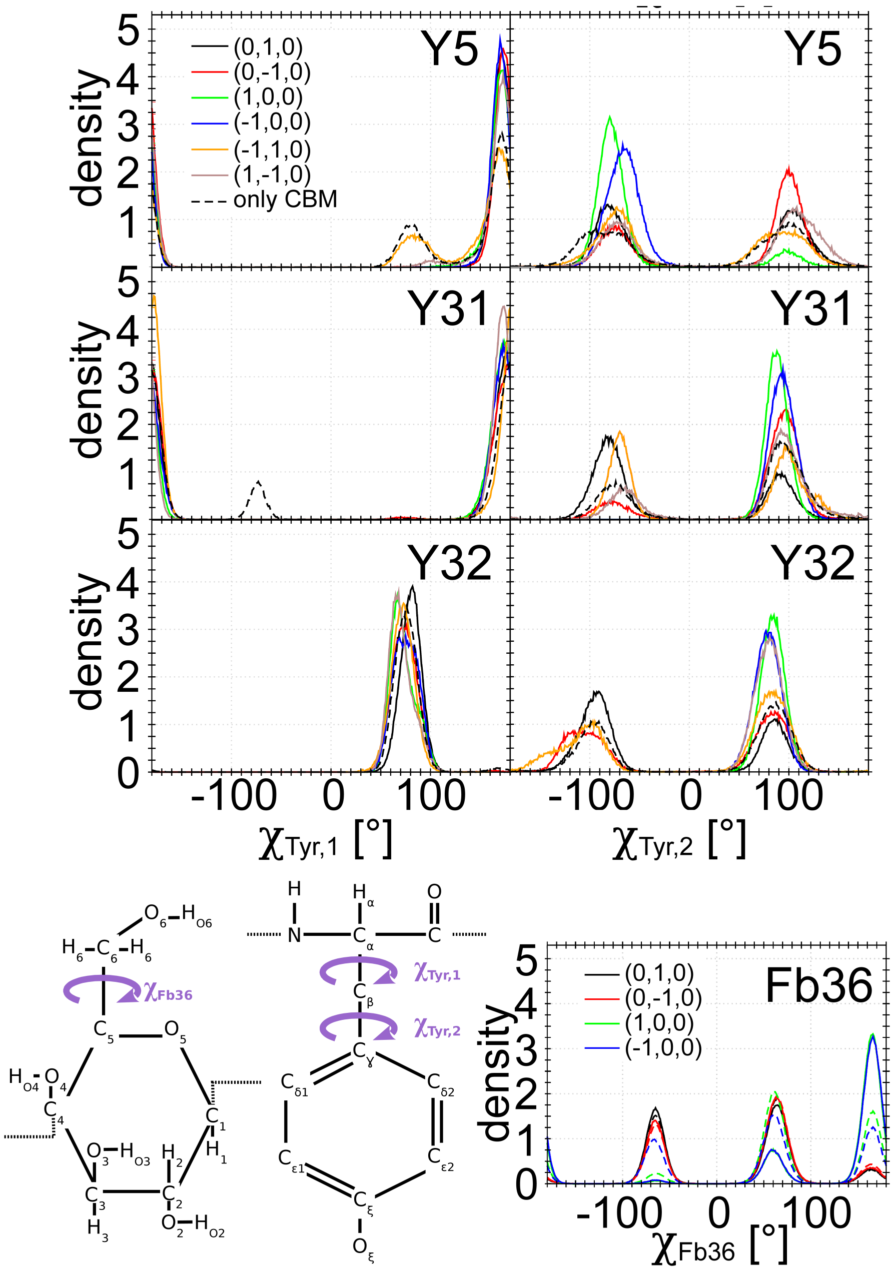


Figure 5.6: (Top) Density distributions of the dihedral angles $\chi_{Tyr,1}$ and $\chi_{Tyr,2}$ of Y5, Y31, and Y32. The dashed lines represent the system containing only the CBM and the solid curves representing the CBM-fiber systems. (Bottom left) Hydroxymethyl and Tyr residue dihedral angles. (Bottom right) Density distributions of the hydroxymethyl dihedral angle χ_{Fb36}.

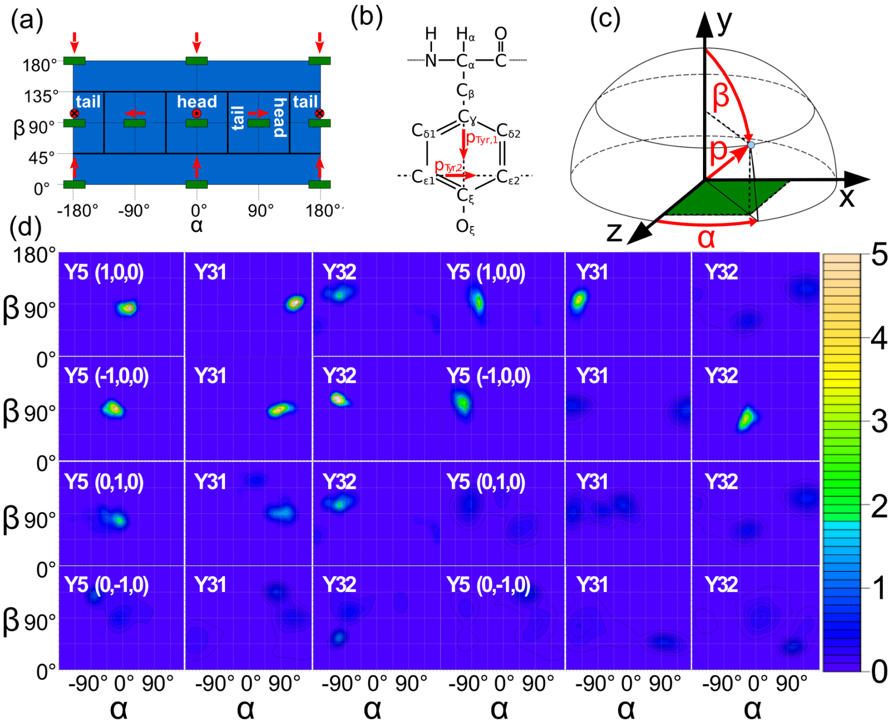


Figure 5.7: (a) Sketch of different orientations of the Tyr ring in the $\alpha-\beta$ map of either $\mathbf{p}_{Tyr,1}$ or $\mathbf{p}_{Tyr,2}$. The head of the vector is denoted with a dot and the tail with a cross. (b) The two vectors $\mathbf{p}_{Tyr,1}$ and $\mathbf{p}_{Tyr,2}$ that define the orientation of a Tyr ring; (c) the components of a vector $\mathbf{p}$ in Cartesian (x, y, z) and spherical polar (r, α, β) coordinates, with α being the azimuth and α the inclination angle; (d) $\alpha-\beta$ maps of the vector $\mathbf{p}_{Tyr,1}$ (columns 1 to 3) and $\mathbf{p}_{Tyr,2}$ (columns 4 to 6) for Y5, Y13, Y31, and Y32 rings on different fiber faces.

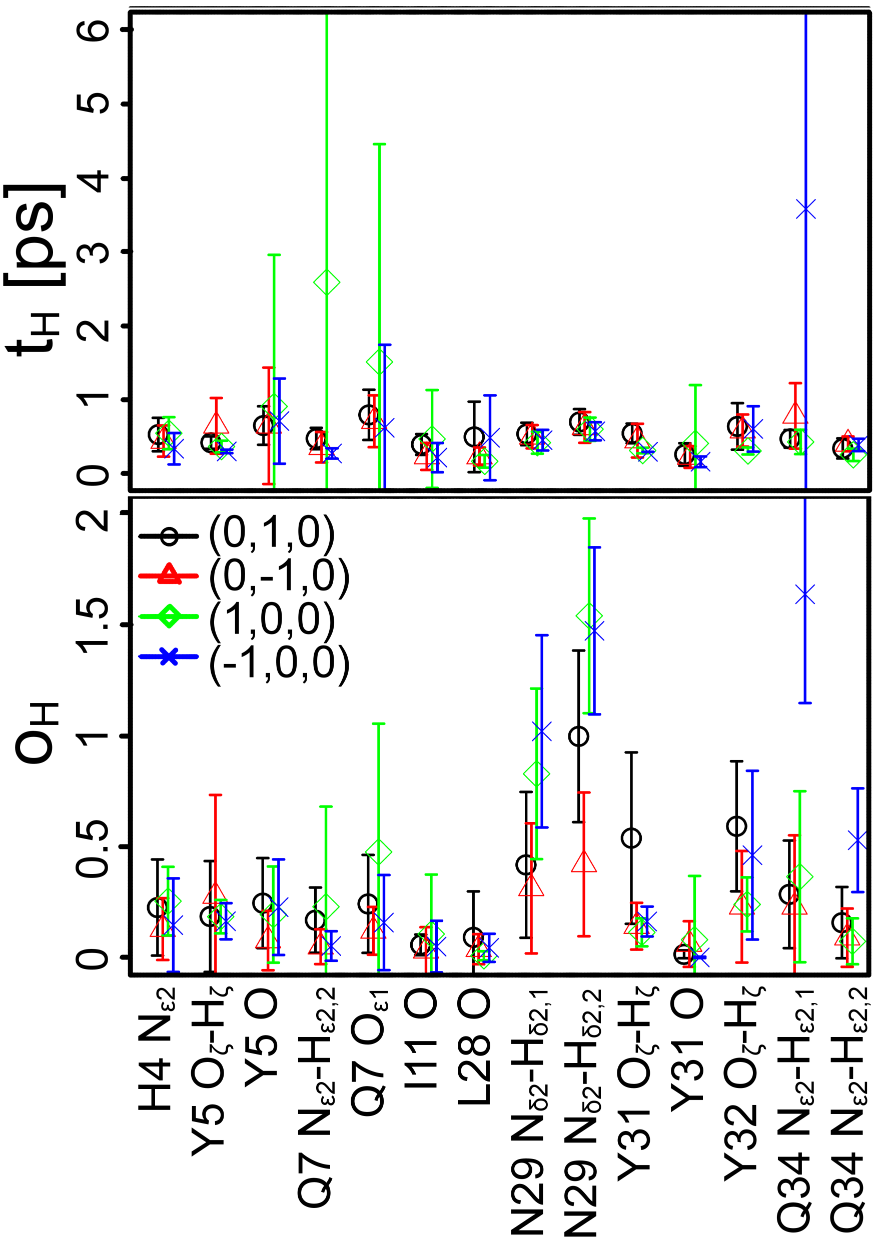


Figure 5.8: Average lifetime, t_H and average occupancy, o_H of representative hydrogen bonds at the interfaces between the CBM and different cellulose fiber faces.

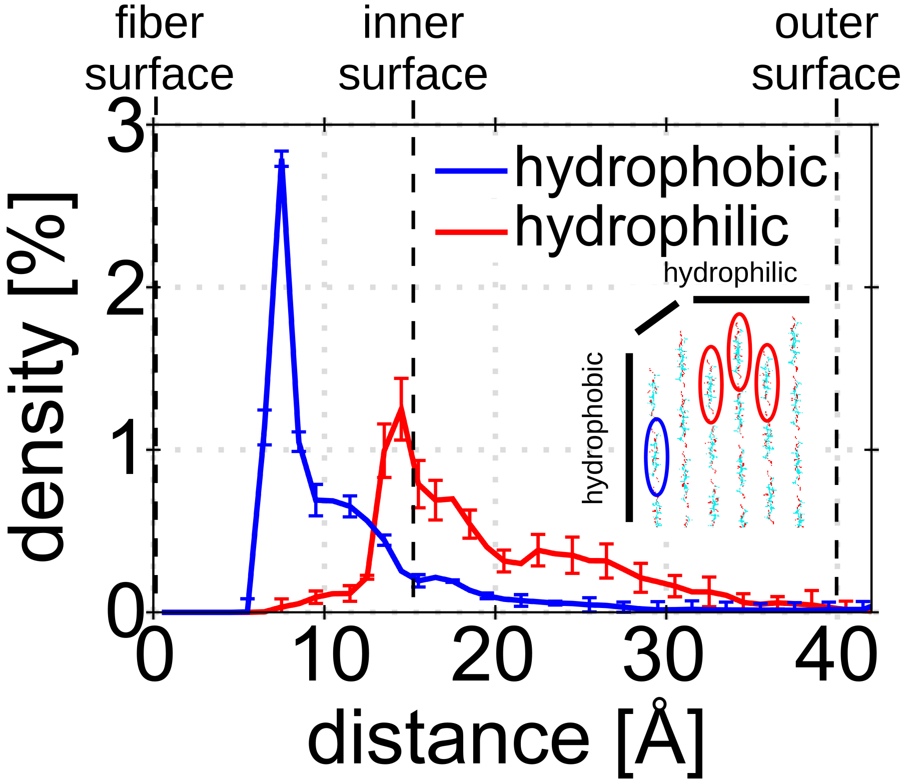


Figure 5.9: Probability distributions of the distances between the CBM hydrophobic patch and the cellulose fiber. The minimal distances between the C_α atoms of Y5, Y31, and Y32 to the fiber surface atoms C_1 are histogrammed and plotted. On both hydrophilic and hydrophobic fiber faces a distinct peak is visible. The CBM cannot approach the hydrophilic fiber faces so close as the hydrophobic ones, likely due in part to the fact that the hydrophilic faces are more "hilly" than the hydrophobic ones (inlay figure).

The science of Nature has already been too long made only a work of the brain and the fancy. It is now high time that it should return to observations.

Robert Hooke 1635-1703

Chapter 6

Role of Cel7A Linker

The research presented in this Chapter is based on the paper entitled "Effect of the Linker Length and Stiffness on the Interaction of Cellulase Cel7A with the Cellulose Iβ Crystal Model" and which is to be submitted.

Prior experimental studies [74, 214] suggest that the binding and enzymatic efficiency of cellulases are affected when the linker is shortened or removed. The role of the Cel7A linker on binding to bacterial microcrystalline cellulose was investigated [74]. Deletion of approximately one third of the linker peptide (residue 18 to 23 in the hinge region, Figure 2.15) resulted in reduced binding capacity of Cel7A but the enzymatic activity on crystalline cellulose was not affected. After deleting the entire linker, Cel7A was found to still bind to cellulose but the degradation activity rate of crystalline cellulose was dramatically reduced. Interaction of the cellulase CenA, which is another cellulase similar to Cel7A, but with different cellulose substrates was also studied in Ref. [214]. The deletion of the 23 amino acid Pro-Thr linker altered the relative orientation of the CD and CBM domains of CenA. The binding of the enzyme to cellulose was not affected but the catalytic efficiency of CenA was reduced. In contrast, the deletion of the Ser-rich linker from xylanase A of *P. fluorescena sep. cellulosa*, which has a structure similar to Cel7A, does not alter the binding properties with avial crystalline cellulose [215]. In a stochastic model in which the CD and CBM of a general cellulase enzyme were described as coupled random walkers it was found that the linker stiffness is an important factor governing the hydrolysis rates [21]. Also, several simulation studies have addressed the role of the Cel7A linker [43, 77, 196]. Ref [196] reported the parameters for a coarse grain forcefield for the isolated Cel7A linker in implicit solvent, using 360 ns of REMD simulations. A 1.5 ns MD study was the first reported simulation on the Cel7A-fiber complex [43]. A 23 ns MD study of the isolated Cel7A linker peptide found that the linker exhibits two stable states [77].

The above experimental, theoretical, and simulation studies indicate that the linker plays a role in the activity of Cel7A on cellulose. The question however remains, what is the in detail role of the Cel7A linker length and stiffness is on the CBM-fiber and CD-fiber interaction with cellulose Iβ. To answer this question I employed a simulation model in which I systematically altered length and stiffness of the linker to investigate the interaction of Cel7A with cellulose Iβ.

In the first stage of the work linker coarse grain parameters are derived from explicit solvent MD simulations with a combined simulation time of >5 μs. I found two different Cel7A linker states, “extended” and “compact”, which I discussed in this study. In the second stage these parameters are used in BD simulations (>45000 trajectories) performed with a combined simulation time of >45 ms. In the BD simulations the linker is modeled as a Hookean spring [216], and I vary the linker length and stiffness to examine their effect on the CBM-fiber and CD-fiber interaction. Based on the BD simulations I propose a linker length and stiffness optimised for both the CBM-fiber and the CD-fiber interactions.

6.1 Simulation and Analysis Details

6.1.1 BD and MD Simulations

The goal of the BD simulation is not to fill the model with undue details but rather to capture the effect of the linker length and flexibility on the Cel7A-fiber interaction. The linker was therefore modeled as a Hookean spring, this assumption being justified by the high fraction of Pro and Thr residues and the linker glycosylation [79,217–219] (Figure 2.15). A high Pro and Gly content are in general indicators for elasticity [217]. Gly, lacking any side chain, provides flexibility. The cyclic side chain of Pro makes the linker peptide stiff and highly restricted in its secondary structure formation. The study [79] showed, that the linker glycosylation can enable an extended conformation. Each coarse grain linker spring model contains less information then a full atomic linker model. To ensure that the results I observe are not a side effect of how the linker is modeled, I study here three different spring models 1 to 3 (Figure 6.4). In model 1 the linker is represented as a single spring, which is attached to the same CBM and CD residues to which the linker was attached. In model 2 the linker is represented also as a single spring, but extending from the center of the CBM to the center of the CD. Therefore the model 2 compared to model 1 has a lower probability that the CBM and CD surface atoms will touch each other (Figure 6.4). Model 3 is a combination of model 1 and 2, the linker is represented by two springs. Linker model 1 is more similar to the full atomic model used in the MD simulations compared to model 2 and 3. The different BD sets are summarized in Table 6.2 and Table 6.3. The CD glycosylation residues were included in the BD simulation.

To obtain reliable linker coarse grain parameters for the BD simulation and analysis ten sets of MD simulations, labelled MD1 to MD3f, were performed (Table 6.1). The sets MD1, MD2a, MD2b, and MD2c contain only the cellulose fiber, or the CBM, or the CD, or the Cel7A, respectively. The sets MD3a to MD3f contain both the Cel7A and the cellulose fiber. For MD3a to MD3f the starting structures, prior to minimization, were created such that the CD and the CBM are at a distance of 3.5 Å from the fiber surface, and this distance was also used in a previous MD study [43]. Each biomolecule was then solvated in a water box with 15 Å to 20 Å solvent padding on all sides. Sodium and chloride counter ions were added to neutralize each system and create an ionic concentration of 100 mM [43].

6.1.2 Spring Stiffness and Equilibrium Length

The linker was modeled as a Hookean spring with a spring stiffness k_{Lk} and a equilibrium length d_{Lk}. Both parameters were calculated from the sets MD2c to MD3f (Table 6.1), by using the potential of mean force (PMF) approach presented in Ref. [196]. The histogram of the end-to-end distance of the linker length d_{LkO} (Figure 2.14) was converted into a relative free energy scale by computing the negative natural logarithm (Figure 6.1). Hooke's law was fitted to the resulting free energy plot to derive k_{Lk} and d_{Lk}.

Name	Biomolecule	Cel7A docking face	Number of trajectories	Length of single trajectory (total simulation time)	Simulation box size	Number of atoms
MD1	36-chain	-	2	240 ns (480ns)	240 × 74 × 68 Å	115537
MD2a	CBM	-	5	150 ns (750 ns)	72 × 57 × 65 Å	24918
MD2b	CD	-	5	150 ns (750 ns)	113 × 94 × 100 Å	101796
MD2c	Cel7A	-	5	¿250 ns (¿1.5 μs)	178 × 97 × 98 Å	158377
MD3a	Cel7A and 36-chain	(1, 0, 0) hydrophobic	5	60 ns (300 ns)	240 × 126 × 120 Å	349975
MD3b	Cel7A and 36-chain	(-1, 0, 0) hydrophobic	5	60 ns (300 ns)	240 × 126 × 120 Å	349975
MD3c	Cel7A and 36-chain	(0, 1, 0) hydrophilic	5	60 ns (300 ns)	240 × 126 × 120 Å	349975
MD3d	Cel7A and 36-chain	(0, -1, 0) hydrophilic	5	60 ns (300 ns)	240 × 108 × 140 Å	349975
MD3e	Cel7A and 36-chain	(-1,1,0) mixed	5	60 ns (300 ns)	240 × 108 × 140 Å	349975
MD3f	Cel7A and 36-chain	(1,-1,0) mixed	5	60 ns (300 ns)	240 × 108 × 140 Å	349975

Table 6.1: Description of different MD simulation sets MD1 to MD3f.

k_{LkO} [$\mathrm{k_BT/Å}$]	0.04	0.35	0.66	1.01	1.5	2	2.5	3	3.5	3.88
$d_{LkO} = 20\,\mathrm{Å}$	-	-	BD1c1	-		-	-	-	-	-
$d_{LkO} = 35\,\mathrm{Å}$	-	-	BD1c2	-	-	-	-	-	-	-
$d_{LkO} = 44.53\,\mathrm{Å}$	BD1a3	-	BD1c3	BD1d3	-	BD1f3	-	-	-	BD1j3
$d_{LkO} = 53\,\mathrm{Å}$	BD1a4	BD1b4	BD1c4	BD1d4	BD1e4	BD1f4	BD1g4	BD1h4	BD1i4	BD1j4
$d_{LkO} = 60.74\,\mathrm{Å}$	BD1a5	-	BD1c5	BD1d5	-	BD1f5	-	-	-	BD1j5
$d_{LkO} = 67\,\mathrm{Å}$	BD1a6	-	BD1c6	BD1d6	-	BD1f6	-	-	-	BD1j6
$d_{LkO} = 74\,\mathrm{Å}$	BD1a7	-	BD1c7	BD1d7	-	BD1f7	-	-	-	BD1j7

k_{LkC} [$\mathrm{k_BT/Å}$]	0.017	0.15	0.28	0.43	0.85	1.28		1.65
$d_{LkC} = 66.81\,\mathrm{Å}$	BD2a	BD2b	BD2c	BD2d	BD2f	BD2h		BD2j

k_{LkO} [$k_BT/\mathring{A}$] k_{LkC} [$k_BT/\mathring{A}$]	0.04 0.017	0.35 0.15	0.66 0.28	1.01 0.43	2.00 0.85	3.00 1.28		3.88 1.65
$d_{LkO} = 44.53\,\mathring{A}; d_{LkC} = 66.81\,\mathring{A}$	BD3a	BD3b	BD3c	BD3d	BD3f	BD3h		BD3j

Table 6.2: In the BD simulation sets different linker equilibrium lengths and stiffness are used, the sets are categorized in a table form. The top, middle, and bottom table summarizes the BD simulation sets for the linker models 1, 2, and 3 respectively.

Table 6.3: The BD simulation set can contain the 36-chain fiber, the CD, and the CBM as Brownian particle (BP). The interaction between the CBM, the CD, and the 36-chain cellulose fiber are modeled using electrostatic (ES) and van der Waals interactions (VDW) terms. If both interaction terms are switched off free diffusion (FD) is observed. Three different models 1 to 3 (Figure 6.4) are used to approximate the linker.

Name	Biomolecule	Nr. of trajectories	Length of single trajectory (total simulation time)	Interaction terms	Spring model
BD0a	36-chain	1	1 μs	FD (1 BP)	-
BD0b	CD	1	1 μs	FD (1 BP)	-
BD0c	CBM	1	1 μs	FD (1 BP)	-
BD0d	Cel7A ($\mathrm{k_{LkO}} = 0.66\,\mathrm{k_BT/Å}$, $\mathrm{d_{LkO}} = 44.53\,\mathrm{Å}$)	1000	350 ns (350 μs)	ES, VDW (2 BP)	1
BD1a3	Cel7A 36-chain ($\mathrm{k_{LkO}} = 0.04\,\mathrm{k_BT/Å}$, $\mathrm{d_{LkO}} = 44.53\,\mathrm{Å}$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	1
BD1a4	Cel7A 36-chain ($\mathrm{k_{LkO}} = 0.04\,\mathrm{k_BT/Å}$, $\mathrm{d_{LkO}} = 53\,\mathrm{Å}$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	1
BD1a5	Cel7A 36-chain ($\mathrm{k_{LkO}} = 0.04\,\mathrm{k_BT/Å}$, $\mathrm{d_{LkO}} = 60.74\,\mathrm{Å}$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	1
BD1a6	Cel7A 36-chain ($\mathrm{k_{LkO}} = 0.04\,\mathrm{k_BT/Å}$, $\mathrm{d_{LkO}} = 67\,\mathrm{Å}$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	1
BD1a7	Cel7A 36-chain ($\mathrm{k_{LkO}} = 0.04\,\mathrm{k_BT/Å}$, $\mathrm{d_{LkO}} = 74\,\mathrm{Å}$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	1
BD1b4	Cel7A 36-chain ($\mathrm{k_{LkO}} = 0.35\,\mathrm{k_BT/Å}$, $\mathrm{d_{LkO}} = 53\,\mathrm{Å}$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	1
BD1c1	Cel7A 36-chain ($\mathrm{k_{LkO}} = 0.66\,\mathrm{k_BT/Å}$, $\mathrm{d_{LkO}} = 20\,\mathrm{Å}$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	1
BD1c2	Cel7A 36-chain ($\mathrm{k_{LkO}} = 0.66\,\mathrm{k_BT/Å}$, $\mathrm{d_{LkO}} = 35\,\mathrm{Å}$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	1
BD1c3	Cel7A 36-chain ($\mathrm{k_{LkO}} = 0.66\,\mathrm{k_BT/Å}$, $\mathrm{d_{LkO}} = 44.53\,\mathrm{Å}$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	1
BD1c4	Cel7A 36-chain ($\mathrm{k_{LkO}} = 0.66\,\mathrm{k_BT/Å}$, $\mathrm{d_{LkO}} = 53\,\mathrm{Å}$)	1000	1μs (1 ms)	ES, VDW (3 BP)	1
BD1c5	Cel7A 36-chain ($\mathrm{k_{LkO}} = 0.66\,\mathrm{k_BT/Å}$, $\mathrm{d_{LkO}} = 60.74\,\mathrm{Å}$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	1

BD1c6	Cel7A 36-chain ($\mathrm{k_{LkO}} = 0.66\,\mathrm{k_B T/\mathring{A}}$, $\mathrm{d_{LkO}} = 67\,\mathrm{\mathring{A}}$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	1
BD1c7	Cel7A 36-chain ($\mathrm{k_{LkO}} = 0.66\,\mathrm{k_B T/\mathring{A}}$, $\mathrm{d_{LkO}} = 74\,\mathrm{\mathring{A}}$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	1
BD1d3	Cel7A 36-chain ($\mathrm{k_{LkO}} = 1.01\,\mathrm{k_B T/\mathring{A}}$, $\mathrm{d_{LkO}} = 44.53\,\mathrm{\mathring{A}}$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	1
BD1d4	Cel7A 36-chain ($\mathrm{k_{LkO}} = 1.01\,\mathrm{k_B T/\mathring{A}}$, $\mathrm{d_{LkO}} = 53\,\mathrm{\mathring{A}}$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	1
BD1d5	Cel7A 36-chain ($\mathrm{k_{LkO}} = 1.01\,\mathrm{k_B T/\mathring{A}}$, $\mathrm{d_{LkO}} = 60.74\,\mathrm{\mathring{A}}$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	1
BD1d6	Cel7A 36-chain ($\mathrm{k_{LkO}} = 1.01\,\mathrm{k_B T/\mathring{A}}$, $\mathrm{d_{LkO}} = 67\,\mathrm{\mathring{A}}$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	1
BD1d7	Cel7A 36-chain ($\mathrm{k_{LkO}} = 1.01\,\mathrm{k_B T/\mathring{A}}$, $\mathrm{d_{LkO}} = 74\,\mathrm{\mathring{A}}$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	1
BD1e4	Cel7A 36-chain ($\mathrm{k_{LkO}} = 1.5\,\mathrm{k_B T/\mathring{A}}$, $\mathrm{d_{LkO}} = 53\,\mathrm{\mathring{A}}$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	1
BD1f3	Cel7A 36-chain ($\mathrm{k_{LkO}} = 2\,\mathrm{k_B T/\mathring{A}}$, $\mathrm{d_{LkO}} = 44.53\,\mathrm{\mathring{A}}$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	1
BD1f4	Cel7A 36-chain ($\mathrm{k_{LkO}} = 2\,\mathrm{k_B T/\mathring{A}}$, $\mathrm{d_{LkO}} = 53\,\mathrm{\mathring{A}}$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	1
BD1f5	Cel7A 36-chain ($\mathrm{k_{LkO}} = 2\,\mathrm{k_B T/\mathring{A}}$, $\mathrm{d_{LkO}} = 60.74\,\mathrm{\mathring{A}}$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	1
BD1f6	Cel7A 36-chain ($\mathrm{k_{LkO}} = 2\,\mathrm{k_B T/\mathring{A}}$, $\mathrm{d_{LkO}} = 67\,\mathrm{\mathring{A}}$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	1
BD1f7	Cel7A 36-chain ($\mathrm{k_{LkO}} = 2\,\mathrm{k_B T/\mathring{A}}$, $\mathrm{d_{LkO}} = 74\,\mathrm{\mathring{A}}$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	1
BD1g4	Cel7A 36-chain ($\mathrm{k_{LkO}} = 2.5\,\mathrm{k_B T/\mathring{A}}$, $\mathrm{d_{LkO}} = 53\,\mathrm{\mathring{A}}$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	1
BD1h4	Cel7A 36-chain ($\mathrm{k_{LkO}} = 3\,\mathrm{k_B T/\mathring{A}}$, $\mathrm{d_{LkO}} = 53\,\mathrm{\mathring{A}}$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	1
BD1i4	Cel7A 36-chain ($\mathrm{k_{LkO}} = 3.5\,\mathrm{k_B T/\mathring{A}}$, $\mathrm{d_{LkO}} = 53\,\mathrm{\mathring{A}}$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	1
BD1j3	Cel7A 36-chain ($\mathrm{k_{LkO}} = 3.88\,\mathrm{k_B T/\mathring{A}}$, $\mathrm{d_{LkO}} = 44.53\,\mathrm{\mathring{A}}$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	1
BD1j4	Cel7A 36-chain ($\mathrm{k_{LkO}} = 3.88\,\mathrm{k_B T/\mathring{A}}$, $\mathrm{d_{LkO}} = 53\,\mathrm{\mathring{A}}$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	1
BD1j5	Cel7A 36-chain ($\mathrm{k_{LkO}} = 3.88\,\mathrm{k_B T/\mathring{A}}$, $\mathrm{d_{LkO}} = 60.74\,\mathrm{\mathring{A}}$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	1
BD1j6	Cel7A 36-chain ($\mathrm{k_{LkO}} = 3.88\,\mathrm{k_B T/\mathring{A}}$, $\mathrm{d_{LkO}} = 67\,\mathrm{\mathring{A}}$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	1
BD1j7	Cel7A 36-chain ($\mathrm{k_{LkO}} = 3.88\,\mathrm{k_B T/\mathring{A}}$, $\mathrm{d_{LkO}} = 74\,\mathrm{\mathring{A}}$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	1
BD2a	Cel7A 36-chain ($\mathrm{k_{LkC}} = 0.017\,\mathrm{k_B T/\mathring{A}}$, $\mathrm{d_{LkO}} = 66.81\,\mathrm{\mathring{A}}$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	2
BD2b	Cel7A 36-chain ($\mathrm{k_{LkC}} = 0.15\,\mathrm{k_B T/\mathring{A}}$, $\mathrm{d_{LkO}} = 66.81\,\mathrm{\mathring{A}}$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	2
BD2c	Cel7A 36-chain ($\mathrm{k_{LkC}} = 0.28\,\mathrm{k_B T/\mathring{A}}$, $\mathrm{d_{LkO}} = 66.81\,\mathrm{\mathring{A}}$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	2

BD2d	Cel7A 36-chain ($k_{LkC} = 0.43\,k_B T/Å$, $d_{LkO} = 66.81\,Å$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	2
BD2f	Cel7A 36-chain ($k_{LkC} = 0.85\,k_B T/Å$, $d_{LkO} = 66.81\,Å$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	2
BD2h	Cel7A 36-chain ($k_{LkC} = 1.28\,k_B T/Å$, $d_{LkO} = 66.81\,Å$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	2
BD2j	Cel7A 36-chain ($k_{LkC} = 1.65\,k_B T/Å$, $d_{LkO} = 66.81\,Å$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	2
BD3a	Cel7A 36-chain ($k_{LkO} = 0.04\,k_B T/Å$, $d_{LkO} = 44.53\,Å$, $k_{LkC} = 0.017\,k_B T/Å$, $d_{LkC} = 66.81\,Å$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	3
BD3b	Cel7A 36-chain ($k_{LkO} = 0.35\,k_B T/Å$, $d_{LkO} = 44.53\,Å$, $k_{LkC} = 0.15\,k_B T/Å$, $d_{LkC} = 66.81\,Å$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	3
BD3c	Cel7A 36-chain ($k_{LkO} = 0.66\,k_B T/Å$, $d_{LkO} = 44.53\,Å$, $k_{LkC} = 0.28\,k_B T/Å$, $d_{LkC} = 66.81\,Å$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	3
BD3d	Cel7A 36-chain ($k_{LkO} = 1.01\,k_B T/Å$, $d_{LkO} = 44.53\,Å$, $k_{LkC} = 0.43\,k_B T/Å$, $d_{LkC} = 66.81\,Å$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	3
BD3f	Cel7A 36-chain ($k_{LkO} = 2\,k_B T/Å$, $d_{LkO} = 44.53\,Å$, $k_{LkC} = 0.85\,k_B T/Å$, $d_{LkC} = 66.81\,Å$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	3
BD3h	Cel7A 36-chain ($k_{LkO} = 3\,k_B T/Å$, $d_{LkO} = 44.53\,Å$, $k_{LkC} = 1.28\,k_B T/Å$, $d_{LkC} = 66.81\,Å$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	3
BD3j	Cel7A 36-chain ($k_{LkO} = 3.88\,k_B T/Å$, $d_{LkO} = 44.53\,Å$, $k_{LkC} = 1.65\,k_B T/Å$, $d_{LkC} = 66.81\,Å$)	1000	1 μs (1 ms)	ES, VDW (3 BP)	3

6.2 Cel7A Extended and Compact State

Here I examine if the Cel7A has one or several different preferred states. First, I look at the different states. Second, I compare the case of Cel7A docked to a cellulose surface with the case of an isolated Cel7A. Third, I look at the driving force for the transition from the extended into the compact state. Fourth, I examine the role of the linker glycosylation. In the last step, I compare my findings with experimental data.

To define different states I examined in the MD simulations (set MD2c) the histogram of the distance $d_{CD,CBM}$ between the C_α atoms of the CBM and CD surface residues (Figure 2.14). The histogram exhibits several minima (Figure 6.2 middle), the position of the first minimum was used to define two states: if $d_{CD,CBM}$ is for the surface side chain atoms $< 7\,\text{Å}$ (for the C_α atoms $< 11\,\text{Å}$) I define Cel7A to be in the compact state, and otherwise it is in the extended state (Figure 6.2 middle and bottom). I have observed in four of the five MD trajectories of set MD2c and all BD trajectories that the free floating Cel7A linker has two different states, here referred to as "extended" and "compact" (Figure 6.3). The MD and BD trajectories were all started in the extended state, the transition of the free floating Cel7A (MD2c and BD0d) to the compact state starts at ≈ 40 ns (for MD see Figure 6.2 top). My MD simulations indicate that the linker glycosylation (Figure 2.15) can not prevent the transition of Cel7A from the extended to the compact state.

Now I examine the Cel7A states for two cases (a) when Cel7A is free floating and (b) when it is docked to a cellulose fiber. The hydrophobic interaction is short range in nature, if the CBM and CD surface C_α atoms in the BD simulations are closer then $< 7\,\text{Å}$ to the hydrophobic fiber face (1, 0, 0) I define Cel7A to be docked to the fiber. I define Cel7A as free floating if the CBM and CD surface C_α atoms have a distance of $> 100\,\text{Å}$ to the fiber surface. The glycosylated linker in theoretical case of being fully extended form has a length of $\approx 99\,\text{Å}$ and the ES interaction is long range, with this two distance cutoffs it is ensured that there is no significant interaction between the enzyme and the fiber. In the Cel7A-fiber complex MD simulation (MD3a to MD3f) the Cel7A surface atoms were in close contact to the fiber surface the transition of Cel7A from the extended into the compact state was not observed in MD3a to MD3f. In the BD simulations the free floating Cel7A prefers to go from the compact into the extended state, when Cel7A docks to the hydrophobic fiber face (1, 0, 0) (Table 6.4). In the BD simulations, once Cel7A escapes from the fiber surface and is free floating, it prefers going back into the compact state (Table 6.4).

Soft linker peptides favors the extended state, with increasing stiffness a larger force is required to get the docked Cel7A back from the compact to the extended state (Table 6.4).
The analysis of the MD simulations (set MD2c) suggests that the driving force for the transition from the extended state into the compact state is hydrophobic as well as polar in nature. The residues Y5, 31, and 32 constitute an hydrophobic patch at the bottom of the CBM, Y13 is slightly buried at the wedge face of the CBM. The solvent accessible solvent area (SASA) of Y5, 13, 31, 32, and the CBM are in the compact state up to $\approx 20\,\%$ smaller compared to the extended state (Table 6.5 top). The visual inspection of the trajectories showed that the hydrophobic area of Y5, 13, 31, and 32 is mainly covered by the CD surface atom and not the linker atoms. To quantify this the C_α -C_α contact map between Y5, 13, 31, 32, and the CD resp. the linker was calculated (Table 6.5 bottom). In the contact map the average distance range for the CD is always $< 9.9\,\text{Å}$ and for the linker it is $> 13.1\,\text{Å}$, which confirms the visual observation. In the contact map for the compact state the polar residues of the CBM are also in close contact with the CD surface. In the BD simulations the internal flexibility of the residues are not modeled, therefore the exact CBM-CD contact residues differ for linker model 1 to 3 (Figure 6.4). The Cel7A compact state is build up of several different binding patterns, I looked at the amino acid patterns in the CBM-CD contact map of the compact mode, for the MD simulations from the set MD2c. Four major binding pattern were observed which have a probability of approx. 27 %, 20 %, 22 %, and 12 % (Table 6.6). The RMSD analysis of the MD simulations (set MD2c) indicates that the CD and CBM packing can change while the Cel7A is in the compact state.

When Cel7A is in the compact mode, the glycosylation of the Cel7A linker peptide (Figure 2.15) prevents the linker atoms from touching the CBM amd CD surface atoms (Figure 6.5 top). The secondary structure of the Cel7A linker peptide was analyzed using the DSPP algorithms [220,221]. A residue i is assigned a turn if there is a hydrogen bond from the O_C to the H_N atom of two neighboring amino acids. It is assigned a bend if the angle between $\{C_\alpha(\mathrm{i}), C_\alpha(\mathrm{i}-2)\}$ and $\{C_\alpha(\mathrm{i}+2), C_\alpha(\mathrm{i})\}$ is $>70°$. The analysis shows, that the steric constraint of the linker peptide glycosylation, induces the formation of turns [220] around the linker peptide residue Ala 10-Pro 11, Arg 13-Thr 14, Gly 18-Pro 19, and Gly 22-Pro 24 (Figure 6.5 and 2.15).
To validate the simulations the radius of gyration R_g of the intact Cel7A was determined according to Guinier and Fournet [222] from the MD simulations (set MD2c) and was compared with the value derived from small angle X-ray scattering

(SAXS) and small angle neutron scattering (SANS). The R_g values of the compact and extended state, calculated from the set Md2a, are 25.5±0.8 Å and 37.5±2.2 Å, respectively (Figure 6.6). They are similar to the experimental values, derived using SAXS [223–225] and SANS [226], which are in the range of 26.1 to 42.7 Å. Because of following reasons, it is difficult with SAXS and SANS experiments to distinguish the extended from the compact state. First, the CBM is too small compared to the CD. Second, the difference between the compact and the extended state is only visible at large q values in the $I(q)$ curve and this region is quite noisy (Figure 6.6 middle). The third reason is, that the sampling in experiments can last several hours to days, Cel7A will change several times between both states. The fourth reason is, that the fitting of the $I(q)$ curve to the experimental derived data is for large q values ambiguous (Figure 6.6 middle). I speculate that the recurrence of a linker peptide in celluloses, might indicate that it is perhaps of evolutionary advantage to join two globular domains, like the CBM and CD, through a unfolded linker peptide. In general the Cel7A encounter process with the fiber takes place in a crowded environment. In the compact state Cel7A is more packed, during a collision of two or more free floating Cel7A, it might be less likely that they get wedged together and the compact Cel7A might have a better chance to reach the cellulose fiber surface. The small angle x-ray study [78] of Cel6A and Cel6B, both have a linker similar to that of Cel7A, concluded that linkers are flexible, and disordered, and can adopt both extended and compact conformations. The study further found that the compact linker were the most frequent and most stable, but that they are able to unwind into longer linker with a relatively low energy cost. The theoretical study [21] suggest that in their stochastic model the maximum hydrolysis rate corresponds to a transition of the linker from a compact to an extended conformation.

6.3 Linker Properties

I explore the hypotheses: Is the linker length d_{LkO} and stiffness k_{LkO} intrinsic properties of the linker. The linker of the free Cel7A (set MD2c) and Cel7A bound to the hydrophilic fiber faces (set MD3c and MD3d) have a k_{LkO} in the range of ≈ 0.66 to $1\,\mathrm{k_B T/Å}$. For the hydrophobic fiber faces k_{LkO} is in the range ≈ 3.87 to $5.11\,\mathrm{k_B T/Å}$ (Table 6.7). The implicit solvent REMD study [196] calculated a k_{LkO} of $0.04\,\mathrm{k_B T/Å}$ for the glycosylated Cel7A linker (no fiber surface present). As a comparison for the order of magnitude, in protein folding experiments spring constants in the range of 0.02 to 0.24 $\mathrm{k_B T/Å}$ are used [227].

Our MD results show that the properties d_{LkO} and k_{LkO} are not intrinsic to the linker but depend on whether explicit or implicit solvent is used, if Cel7A is docked to fiber surface or not, on the hydrophobicity of the fiber surface and whether a complete Cel7A or only an isolated linker is studied (column d_{LkO} and k_{LkO} in Table 6.7, Figure 6.1). An isolated linker will behave differently from a linker in an intact Cel7A, because the CD and CBM domains are comparably much bulky which exert their influence on the linker dynamics, i.e., when the linker is is coupled to CD and CBM the motion of the linker no longer resembled regular polymer motion but is dominated by the protein domain motion [228–230]. When Cel7A was bound to a cellulose fiber the hydrophobic compared to the hydrophilic fiber faces makes the linker a very hard spring (Table 6.7). This can be explained by the fact that the CBM hydrophobic patch makes Cel7A bound stronger to the hydrophobic then the hydrophobic fiber faces.

6.4 Influence Linker Length and Stiffness on CBM-Fiber Interaction

The linker couples the CD to the CBM, therefore the linker length and stiffness might effect the CBM-fiber interaction. In order to account for their effect on the CBM-fiber interaction, systematically changed both parameters in the BD simulations.

Effect of Linker Length. To study the effect of the linker length on the CBM-fiber interaction I looked at the fraction of time, the enzyme is in the extended state. I compared the case of the enzyme when it is docked to the cellulose fiber and with a free floating enzyme. The linker length was gradually increased the enzyme stays an increased fraction of time in the extended state (Table 6.4). This suggests that the linker length regulates the transition probabilities between the compact and the extended state. A short linker causes the enzyme to get trapped in the compact state, an enzyme with a long linker on the other hand favors the enzyme being in the extended state (Table 6.4). The density map analysis shows that in case of a long compared to a short linker, the CBM can more often bind the enzyme to the hydrophobic fiber face (Figure 6.7, Table 6.4). With increasing linker length the CBM binds more often to the hydrophobic fiber face (1, 0, 0). The CBM can bind with the hydrophobic residues Y5, 31, and 32 either to the fiber or to the CD. A long linker might enable enough separation between the CD and the CBM (Table 6.4). In summary, Figure 6.7 d indicates an increasing thermodynamic CBM-fiber binding preference with increasing linker length.

Effect of Linker Stiffness. Now I analyze the effect of the linker stiffness on the CBM-fiber interaction. I look at the density map bins close to the hydrophobic fiber face (1, 0, 0) and (-1, 0, 0), respectively. I calculated the sum of the 50 and 1600 bins which are most populated by the CBM (which I denote as h_{50} and h_{1600}, respectively, Figure 4.14). The BD trajectories are all started above the fiber face (1, 0, 0) (Figure 4.6 top left, Figure 6.8 a and c), the fiber face (-1, 0, 0) has the farthest distance from the starting position (Figure 6.8 b and d). If the linker gets softer h_{50} and h_{160}, respectively, are increasing (Figure 6.8 b and d). This suggests that for a soft linker the CBM is more mobile and can visit more often the hotspot at (-1, 0, 0) compared to a stiff linker (Figure 6.8 right two figures). The linker stiffness influences the CBM-fiber interaction via the CBM mobility. The CD has a higher mass compared to the CBM. In the theoretical case of a very loose linker the CBM motion is completely independent of the CD motion, the CBM can then freely sample the fiber and bind to it. In the other extreme case of a very stiff linker the CBM and CD act as a single body and the CD slows down the motion of the CBM.

6.5 Influence of Linker Length and Stiffness on CD-Fiber Interaction

After the CBM is docked to the cellulose fiber, the next step would be to thread a lose cellulose chain into the CD tunnel. The exact role of the linker on the CD-fiber interaction is not understood in detail [21]. In particular, does the CD bind to the fiber? When the CBM is docked to a fiber face, does the CD stay on the same fiber face like the CBM or does the CD prefer one of the two neighboring fiber faces? During the CBM docking, does the CBM push the CD via the linker peptide or is the CD pulled by the CBM?

Relative CD Position after CBM-Fiber Docking. There are different hypothesis [55] which speculate that the CBM uses its bottom or wedge surface to thread a lose cellulose chain into CD tunnel after the CBM is docked to the fiber. The CD-CBM interaction, during the threading, process might be eased if the CD and CBM stay on the same fiber face. When the CBM is docked to the hydrophobic fiber face (1, 0, 0) I observed in the BD trajectories that the CD was located at the same fiber face as the CBM as well as the two neighboring fiber faces (0, 1, 0) and (1, -1, 0). To quantify this I calculated the probability p_{sf} that the CD and CBM stay on the same fiber face (Table 6.8). All three fiber faces show a similar probability (Table 6.8), this indicates that the CD has no significant preference to stay with the

CBM on the (1, 0, 0) fiber face. The CD prefers the two neighboring fiber faces similar to the CBM docking fiber face.

Does CD Bind to Fiber. The CBM can bind via its hydrophobic patch to the hydrophobic fiber faces, it is unclear if the CD prefers to bind to any specific fiber face. The comparison of the CBM and CD density maps show that the CBM hotspots are located mainly at the hydrophobic fiber face (1, 0, 0), the CD density map hotspots are less pronounced compared to the ones of the CBM (data for simulation set BD2c3 shown in Figure 4.14). This indicates that the CD compared to the CBM is less restricted to a fiber position. The h_{50} and h_{1600} analysis for the different fiber faces does not show any clear tendency for the effect of the linker length or stiffness of the CD-fiber interaction.

6.6 Optimal Linker Length and Stiffness

The cellulase linker length and glycosylation pattern, which can effect the linker stiffness, vary a lot [231]. A long and soft linker can be favorable for the CBM-fiber interaction, allowing the CBM to sample the fiber independent of the CD, but is unfavorable for the CD-fiber interaction because the CD is not kept close to the hydrophobic fiber face. A short and stiff linker on the other hand keeps the CD and CBM on the same fiber face which might improve the threading of the lose cellulose chain into the CD tunnel, but this slows down the CBM mobility and can hinder the CBM in scanning the fiber surface. With the BD results I determine the optimal linker length d_{LkO}^{Opt} and stiffness k_{LkO}^{Opt}, for which both the CBM-fiber and CD-fiber interactions are optimized. For the optimization I use as a first criteria that the CBM has an high mobility, approximated by an high h_{50} for fiber face (-1, 0, 0) (Figure 6.8), and as a second criteria that the CD stays close to the hydrophobic fiber face, approximated by an high p_{sf} (Table 6.8). To compare h_{50} with p_{sf} both are converted to a new scale s_h and s_{sf} by mapping them linear to the range of 0 to 1, with 0 being a not optimal and 1 an optimal value:

$$s_h(h_{50}) = \frac{max(h_{50}) - h_{50}}{max(h_{50}) - min(h_{50})}\,, \tag{6.1}$$

$$s_{sf}(p_{sf}) = \frac{max(p_{sf}) - p_{sf}}{max(p_{sf}) - min(p_{sf})}\,. \tag{6.2}$$

The discrete values d_{LkO} and k_{LkO} from the BD sets are interpolated with splines [232,233] to get intermediate values. These optimal linker parameters d_{LkO}^{Opt} and k_{LkO}^{Opt}

will maximize s_h as well as s_{sf}. Following Occam's razor principle [234,235], the model with the fewest assumptions should be chosen, I weight s_h and s_{sf} with the factor ω resp. 1-ω which gives the model $s(k_{LkO}, d_{LkO}) = [\omega \cdot s_h(k_{LkO}, d_{LkO}) + (1-\omega) \cdot s_{sf}(k_{LkO}, d_{LkO})]$, with ω being in the range 0 to 1. The weight ω=0.5 means that both criteria have equal importance in the optimization. For this case, the maximum of $s(k_{LkO}, d_{LkO})$ gives $k_{LkO}^{Opt} = 0.49\, k_B T/\AA$ and $d_{LkO}^{Opt} = 69.1\, \AA$ (Figure 6.9), this is slightly larger then the value for the isolated Cel7A which nature has chosen (set MD2c, Table 6.7). For this value the CBM has a high mobility to fast sample the fiber surface and at the same time the CD has a high probability to stay close to the hydrophobic fiber face. Experimentalist might optimize the Cel7A by adjusting the linker length and glycosylation pattern.

6.7 Summary

Previous studies [74, 212, 214] indicate that the linker might be essential for the cellulase-fiber interaction. In this study I systematically investigated the role of the Cel7A linker length and stiffness.

The new contributions which this study provides are, that the equilibration length and stiffness are not intrinsic properties of the linker. They depend whether implicit or explicit solvent is used, whether the complete Cel7A or only the isolated linker is studied and the hydrophobicity of the fiber surface to which Cel7A is docked.

MD and BD simulations show that the linker of a free floating Cel7A can not guarantee a spatial distance between the CD and CBM. I showed that the free floating Cel7A linker has two different states which I call extended and compact. The Cel7A packing in the compact state is hydrophobic as well as polar in nature.

The linker length regulates the transition between the extended and compact linker states of the enzyme. The linker stiffness influences the CBM mobility, I furthermore quantified the effect of the linker stiffness on the CBM mobility. The properties and dynamic of the free Cel7A are compared with the one of the Cel7A-fiber complex, in particular the influence of the fiber surface hydrophobicity on the Cel7A dynamics. I extended the set of the available coarse grain parameters, in particular the linker spring stiffness and equilibration length, of Cel7A. I also provide coarse grain parameters in dependence of the fiber surface hydrophobicity and the effect of explicit solvent. These parameters will provide more realistic models for the linker in coarse grain simulations of the Cel7A-fiber complex. I determined the optimal linker length d_{LkO} and stiffness k_{LkO} for the CBM-fiber and CD-fiber interaction.

In summary my study shows that the linker is not an accessory part of Cel7A. The linker length and stiffness are not random properties but that they have a significant effect on the Cel7A-fiber interaction. The new insights obtained here can assist in designing Cel7A with mutated linker that have improved hydrolysis properties, without having to alter the CBM or CD domains.

Spring model	Set		Fraction time enzyme is in "extended" state	
			enzyme docked < 7 Å	enzyme free floating > 100 Å
1	BD1a3		44.2 %±0.1 %	1.0 %
1	BD1a4		45.5 %±0.2 %	1.1 %
1	BD1a5	d_{LkO} ↓	50.0 %±0.2 %	1.3 %
1	BD1a6		53.4 %±0.1 %	1.5 %
1	BD1a7		56.8 %±0.1 %	1.6 %
1	BD1c1		21.1 %±1.1 %	0.5 %
1	BD1c2		32.5 %±1.6 %	0.7 %
1	BD1c3		39.4 %±0.1 %	0.8 %
1	BD1c4	d_{LkO} ↓	42.8 %±1.6 %	1.0 %
1	BD1c5		46.3 %±2.8 %	1.1 %
1	BD1c6		48.1 %±1.8 %	1.2 %
1	BD1c7		52.3 %±1.4 %	1.3 %
1	BD1j3		39.9 %±0.9 %	0.9 %
1	BD1j4		43.8 %±1 %	1.0 %
1	BD1j5	d_{LkO} ↓	45.6 %±0.9 %	1.2 %
1	BD1j6		46.3 %±1 %	1.3 %
1	BD1j7		50.9 %±0.4 %	1.5 %
1	BD1a4		45.5 %±0.2 %	1.1 %
1	BD1c4		42.8 %±1.6 %	1.0 %
1	BD1d4		43.7 %±3.3 %	1.0 %
1	BD1e4		41.7 %±0.8 %	1.0 %
1	BD1f4	k_{LkO} ↓	41.5 %±1.5 %	1.0 %
1	BD1g4		42.9 %±2.3 %	1.0 %
1	BD1h4		42.1 %±0.7 %	1.0 %
1	BD1i4		43.4 %±0.5 %	1.0 %
1	BD1j4		43.8 %±1.0 %	1.0 %
2	BD2a		58.6 %±0.2 %	1.4 %
2	BD2b		54.5 %±1.4 %	1.2 %
2	BD2c		54.1 %±0.6 %	1.1 %
2	BD2d	k_{LkC} ↓	53.3 %±2.2 %	1.1 %
2	BD2f		50.8 %±1.0 %	1.1 %
2	BD2h		53.8 %±1.3 %	1.1 %
2	BD2j		53.2 %±0.8 %	1.1 %
3	BD3a		42.4 %±0.8 %	1.0 %
3	BD3b		38.9 %±0.6 %	0.8 %
3	BD3c	k_{LkO} ↓	40.5 %±0.5 %	0.8 %
3	BD3d	k_{LkC} ↓	38.7 %±1.4 %	0.8 %
3	BD3f		37.6 %±0.6 %	0.8 %
3	BD3h		36.9 %±0.1 %	0.9 %
3	BD3j		37.9 %±0.6 %	0.8 %

Table 6.4: Cel7A has two states "extended" and "compact" (Figure 6.3). The arrow in the second column indicates in which direction d_{LkO}, k_{LkO}, and k_{LkC} are increasing. The enzyme has higher probability to be in the extended state when it is docked to the fiber compared to the case when it is free floating. With increasing linker length d_{LkO} the enzyme has an higher probability to be in the extended state.

	Average solvent accessible surface area [Å^2]		
	extended	compact	difference
Y5	147.71±10.76	123.39±36.88	19.7 %
Y13	70.15±9.34	67.83±13.38	3.4 %
Y31	141.05±13.94	117.62±35.75	19.9 %
Y32	81.87±10.22	55.63±31.31	20.8 %
Y5, 13, 31, and 32	440.19±24.34	364.48±79.18	20.8 %
entire CBM	2447.57±62.63	2258.3±105.42	8.4 %

CBM	CD [Å]	linker [Å]
Y5	≤9.9	≥14.2
Y13	≤8.9	≥12.1
Y31	≤9.3	≥13.2
Y32	≤9.7	≥13.2

Table 6.5: (Top) Average solvent accessible surface area of the CBM hydrophobic residues (Y5, 13, 31, and 32) and the entire CBM in simulation set MD2c. (Bottom) Average distance range between the CBM Y5, 13, 31, and 31 C_α and the CD resp. linker C_α atoms in simulation set MD2c, distances larger then 15 Å are ignored. The table shows that the hydrophobic area of Y5, 13, 31, and 32 are covered by the CD surface atoms and not the linker surface atoms.

CBM	CD residues			
residues	pattern 1	pattern 2	pattern 3	pattern 4
Y5	-	C4, T5	-	-
Y31	T5, L6, S8, S9	-	-	-
Y32	C4, T5, Q7, S8	-	-	-
T1	-	-	S21, G22	-
G9	S87	-	-	S87
I11	Q7	-	-	-
V18	-	-	S21, G22	-
C19	-	-	G22	-
A20	-	C4, T5	G22, G23	P432
S21	-	C4, T5	G22, G23	N431, P432, S433
P30	A100, Q101	-	-	-
probability	≈27 %	≈20 %	≈22 %	≈12 %

Table 6.6: 11: In the MD simulation of set MD2c four different binding patterns of the Cel7A compact state were observed. The relevant CD and CBM amino acids of the contact map are presented. For each compact state pattern the probability is given.

MD simulation set	d_{LkO} [Å]	k_{LkO} [$\mathrm{k_BT}$/Å]	d_{LkC} [Å]	k_{LkC} [$\mathrm{k_BT}$/Å]
only glyco. Cel7A (MD2c)	44.53 ± 7.77	0.664 ± 0.063	66.81 ± 13.52	0.278 ± 0.025
Cel7A-fiber (1, 0, 0) hydrophobic (MD3a)	60.74 ± 2.63	3.883 ± 0.086	-	-
Cel7A-fiber (-1, 0, 0) hydrophobic (MD3b)	60.89 ± 2.82	5.126 ± 0.174	-	-
Cel7A-fiber (0, 1, 0) hydrophilic (MD3c)	57.97 ± 7.97	1.013 ± 0.19	-	-
Cel7A-fiber (0, -1, 0) hydrophilic (MD3d)	59.5 ± 3.16	0.892 ± 0.123	-	-
Cel7A-fiber (-1,1,0) mixed (MD3e)	57.34 ± 3.62	2.253 ± 0.173	-	-
Cel7A-fiber (1,-1,0) mixed (MD3f)	56.57 ± 4.06	2.698 ± 0.186	-	-
[196] (isolated nonglyco. linker of Cel7A, implicit REMD)	37	0.05	-	-
[196] (isolated glyco. linker of Cel7A, implicit REMD)	53	0.04	-	-

Table 6.7: Equilibrium length d_{Lk} and spring stiffness k_{Lk} for the linker, obtained from MD simulations. For a visualization of the spring parameters d_{LkO}, k_{LkO}, d_{LkC}, and k_{LkC} view spring models in Figure 6.4.

Spring model	Set		p_{sf}
1	BD1a4	$k_{LkO}\downarrow$	35.5 %
1	BD1b4		33.9 %
1	BD1c2		36.3 %
1	BD1d4		35.7 %
1	BD1e4		33.5 %
1	BD1f4		37.5 %
1	BD1g4		32.1 %
1	BD1h4		32.6 %
1	BD1i4		37.1 %
1	BD1j4		37.5 %
1	BD1a3	$d_{LkO}\downarrow$	30.9 %
1	BD1a4		35.5 %
1	BD1a5		35.1 %
1	BD1a6		32.4 %
1	BD1a7		36.2 %
1	BD1c1	$d_{LkO}\downarrow$	35.7 %
1	BD1c2		36.5 %
1	BD1c3		34.9 %
1	BD1c4		36.3 %
1	BD1c5		35.7 %
1	BD1c6		37 %
1	BD1c7		31.3 %
1	BD1d3	$d_{LkO}\downarrow$	36.7 %
1	BD1d4		35.7 %
1	BD1d5		36.7 %
1	BD1d6		34.5 %
1	BD1d7		35.4 %
1	BD1f3	$d_{LkO}\downarrow$	38.9 %
1	BD1f4		37.5 %
1	BD1f5		33.1 %
1	BD1f6		33.6 %
1	BD1f7		33.7 %
1	BD1j3	$d_{LkO}\downarrow$	38.3 %
1	BD1j4		37.48 %
1	BD1j5		32.2 %
1	BD1j6		32.4 %
1	BD1j7		31.9 %

Spring Model	Set		p_{sf}
2	BD2a	$k_{LkC}\downarrow$	27.9 %
2	BD2b		32.2 %
2	BD2c		33.7 %
2	BD2d		36.4 %
2	BD2f		31.5 %
2	BD2h		32.4 %
2	BD2j		30.3 %
3	BD3a	$k_{LkO}\downarrow$ $k_{LkC}\downarrow$	35.2 %
3	BD3b		36.2 %
3	BD3c		38.5 %
3	BD3d		36.6 %
3	BD3f		38.1 %
3	BD3h		36,7 %
3	BD3j		37.9 %

Table 6.8: When CBM is docked to the hydrophobic fiber face (1, 0, 0) I calculated if the CD is closer to the (1, 0, 0) or the two neighboring fiber faces (0, 1, 0) and (1, -1, 0) (Figure 4.6 top left). p_{sf} gives the probability to find the CD on the same fiber face as the CBM. The results show that the CD has no preference to stay on the same fiber face as the CBM but rather diffuses away to the two neighboring fiber faces.

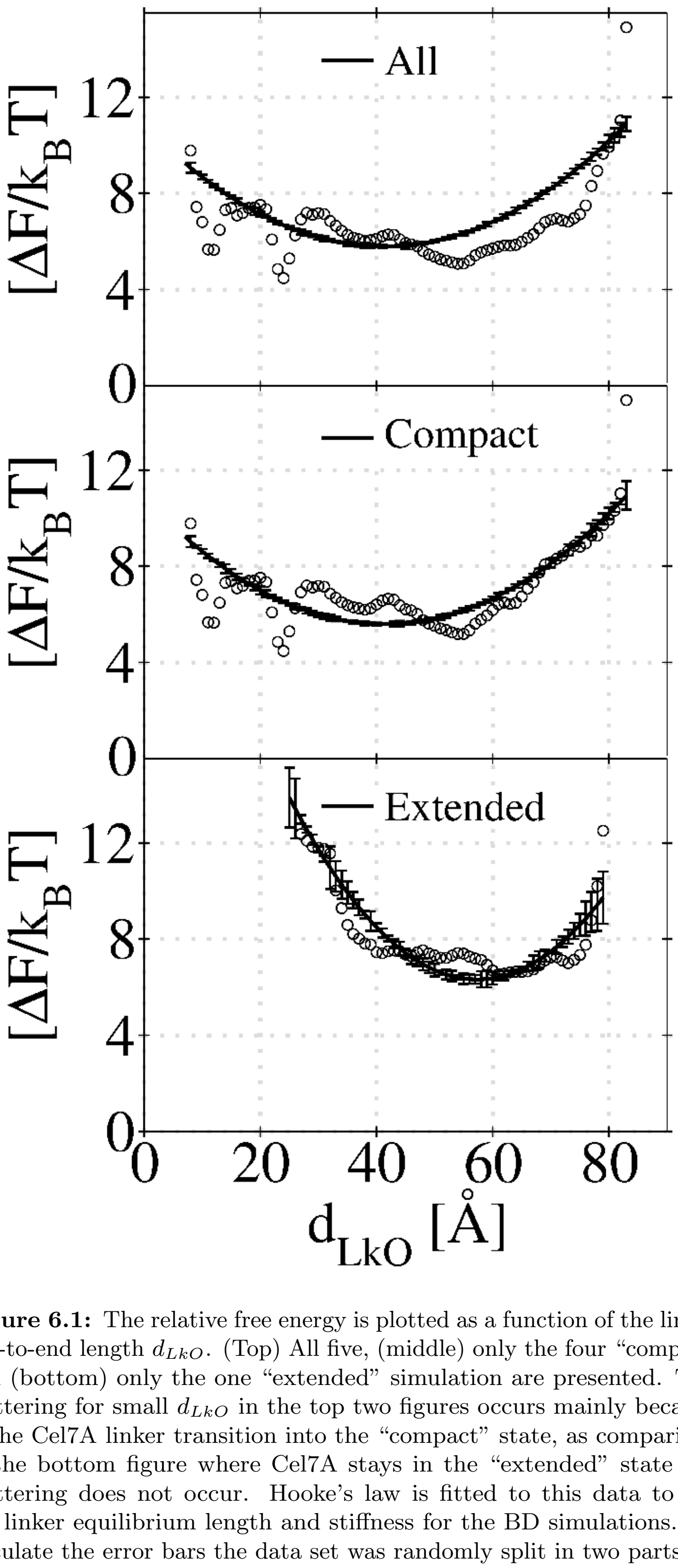


Figure 6.1: The relative free energy is plotted as a function of the linker end-to-end length d_{LkO}. (Top) All five, (middle) only the four “compact” and (bottom) only the one “extended” simulation are presented. The scattering for small d_{LkO} in the top two figures occurs mainly because of the Cel7A linker transition into the “compact” state, as comparison in the bottom figure where Cel7A stays in the “extended” state the scattering does not occur. Hooke’s law is fitted to this data to get the linker equilibrium length and stiffness for the BD simulations. To calculate the error bars the data set was randomly split in two parts, to each data set Hooke’s law was fitted, the difference in the fit gives the error bar.

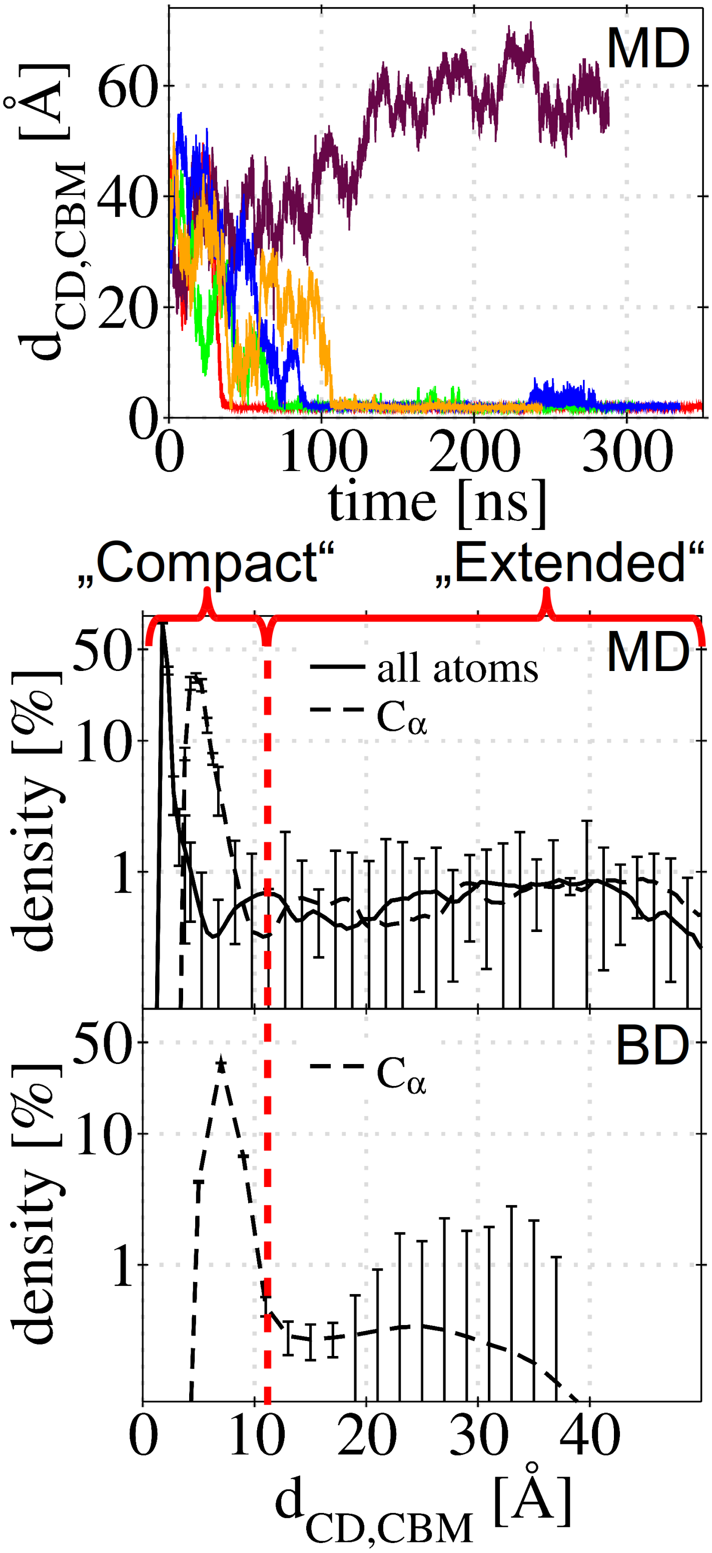


Figure 6.2: (Top) Time evolution of minimal distance $d_{CD,CBM}$ between the CD and the CBM surface atoms in the MD set MD2c. Each of the five Cel7A trajectories is shown in a different color. (Middle and Bottom) The histogram of $d_{CD,CBM}$ for the MD (MD2c) and BD (BD0d) simulations shows that two distinct states “compact” and “extended” exist. The first minima in the histogram of $d_{CD,CBM}$ for the C_α atoms from the MD simulation is used for the formal definition of two Cel7A linker states.

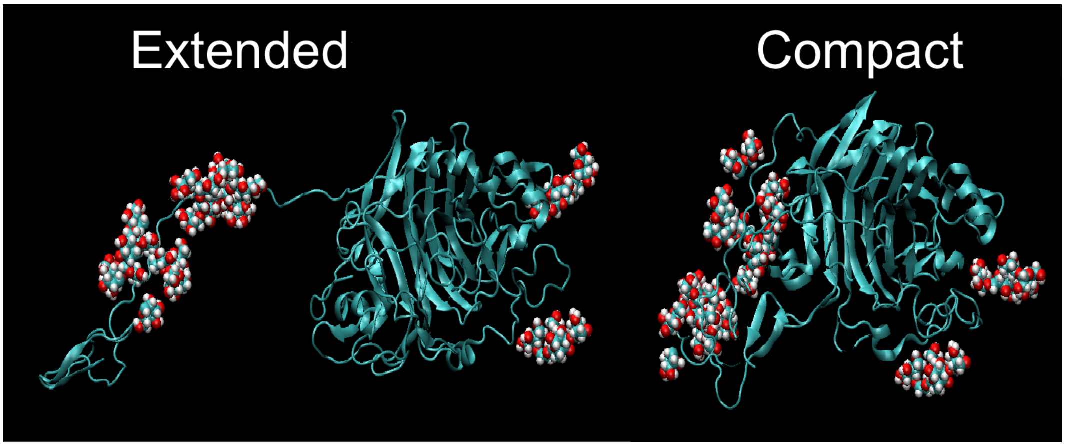


Figure 6.3: Visualization of the two Cel7A linker states “extended” and “compact”.

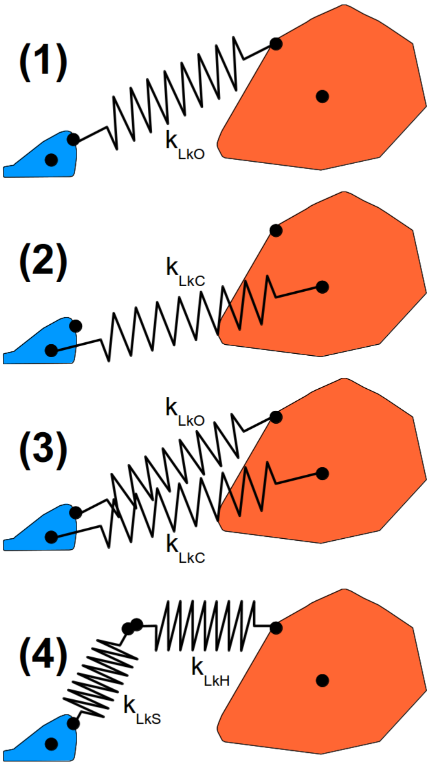


Figure 6.4: Schematic view of the different spring models for the Cel7A linker. In the BD simulations the linker can be modeled in four different ways. (1) The linker is modeled as a single Hookean spring which is attached to the CBM and CD residue to which the linker was attached. (2) Or as a single spring which goes from the center of the CBM to the center of the CD. (3) Or via two springs, by combining model 2 and 3. (4) The hinge and the stiff part of the linker can be modeled as two serial springs.

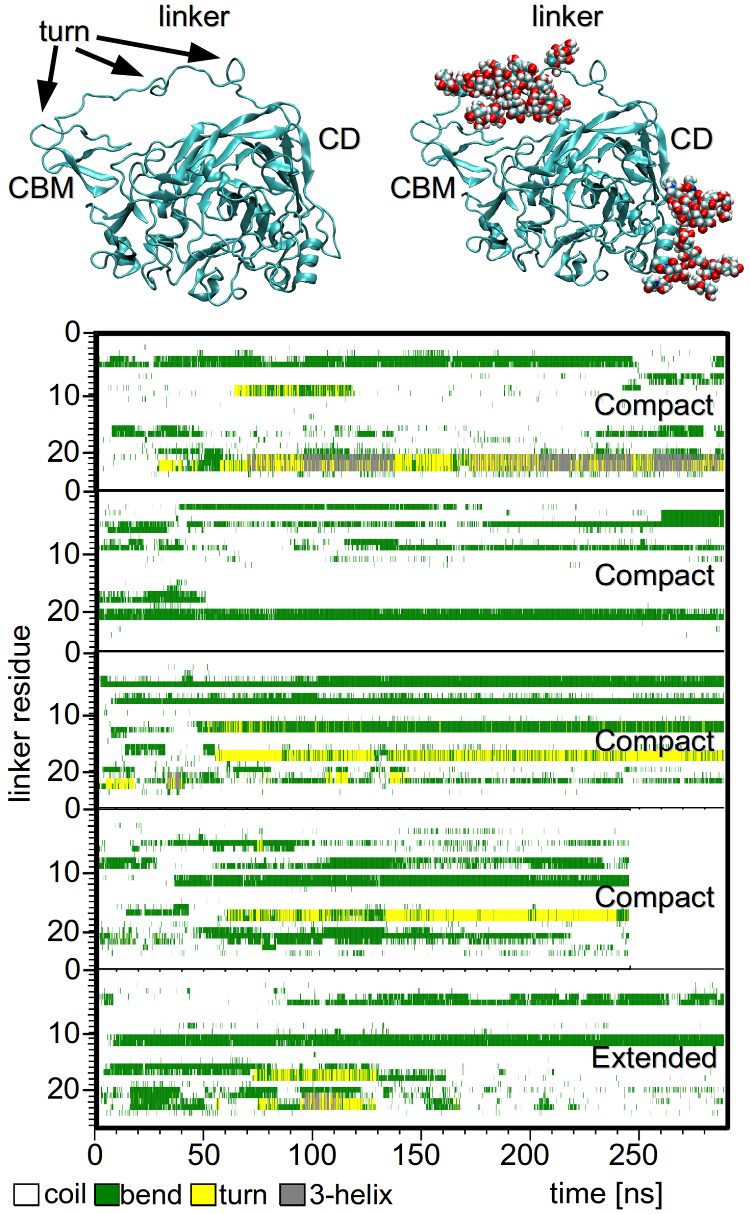


Figure 6.5: (Top left) Schematic view of the linker in the Cel7A compact state without the linker glycosylation (top left) and with the linker glycosylation (top right). (Bottom 5 rows) For each of the five trajectories from set MD2c the secondary structure of the Cel7A linker peptide is visualized. The linker glycosylation prevents the linker atoms from touching the CBM and CD surface atoms.

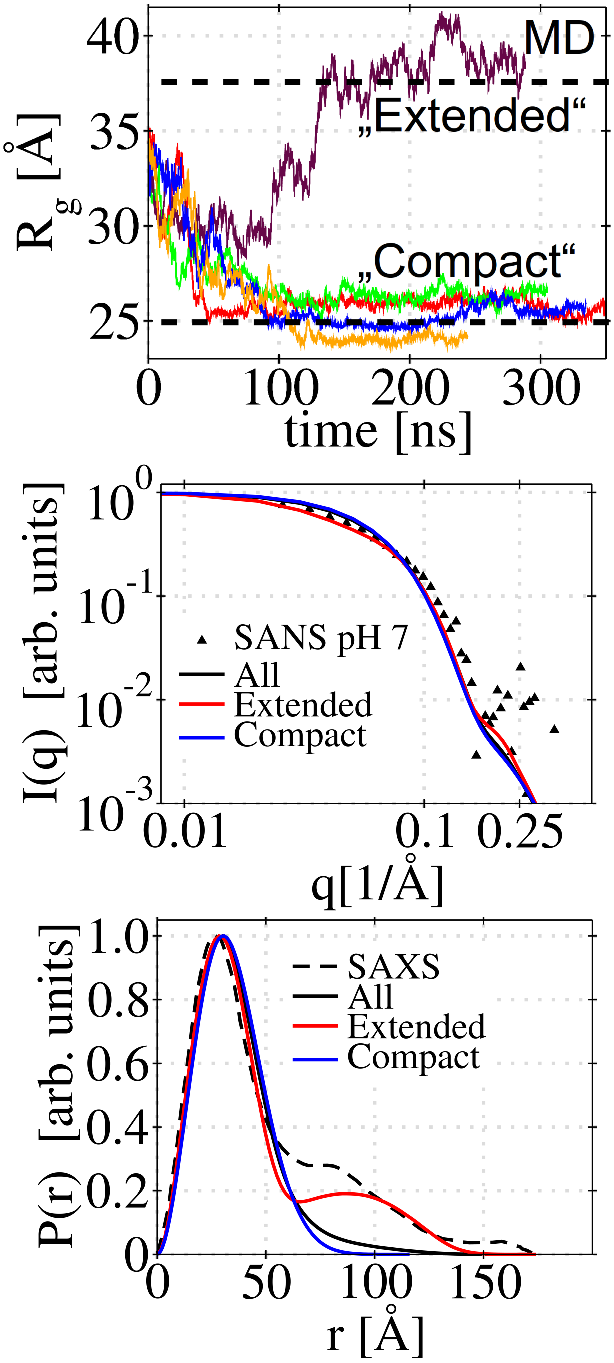


Figure 6.6: (Top) Radius of gyration R_g of the intact Cel7A calculated from MD simulations (set MD2c). The simulations were started in the extended state, four out of five MD trajectories went into the compact state and one stayed in the extended state. The average R_g (first 100 ns neglected) of the compact and extended state are visualized via a dashed line.(Middle) $I(q)$ scattering curve, experimental SANS data from Ref. [226] is used. (Bottom) Distance distribution function $P(r)$ calculated from MD simulations (set MD2c) and from SAXS data [223].

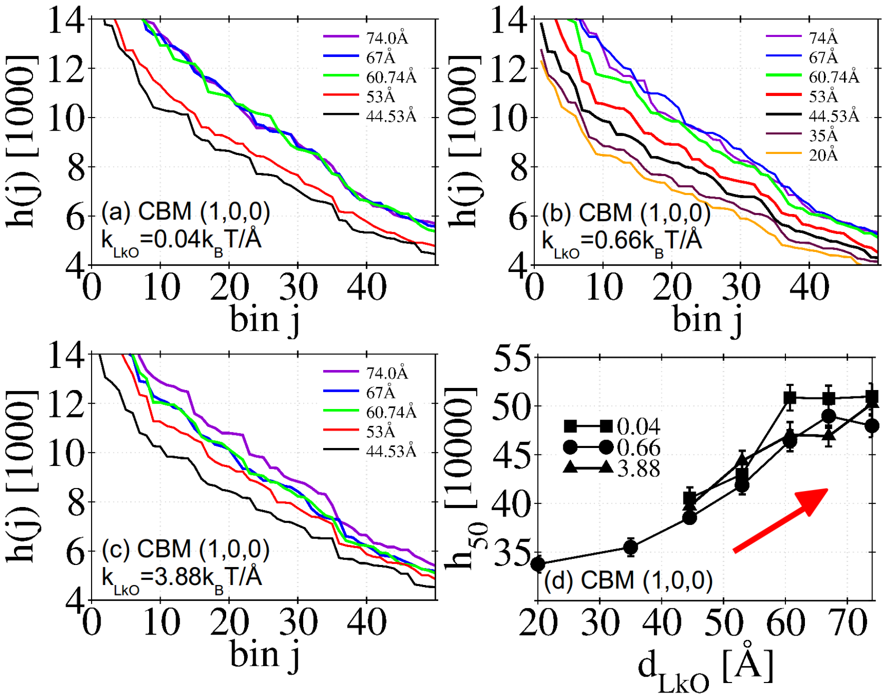


Figure 6.7: In the BD density maps (e.g. Figure 4.14) I analyzed how often the center of the CBM visits the bins close to the hydrophobic fiber face (1, 0, 0). The 50 most populated bins are sorted in decreasing order. In the new sorted set $h(j)$, the index j=1 refers to the most populated bin and j=50 to the least populated bin. In sub figure a, b, and c the bin j is plotted against $h(j)$ for different linker lengths d_{LkO} and stiffness k_{LkO}. For a fixed spring stiffness the curve $h(j)$ for longer linker lengths are always above the curves with a smaller linker length. In sub figure d the sum of the 50 bins, denoted as h_{50} is plotted as a function of the linker length d_{LkO}. Sub figure (a) to (d) indicate that with increasing linker length also the thermodynamic binding preference of the CBM to the hydrophobic fiber face (1, 0, 0) also increases.

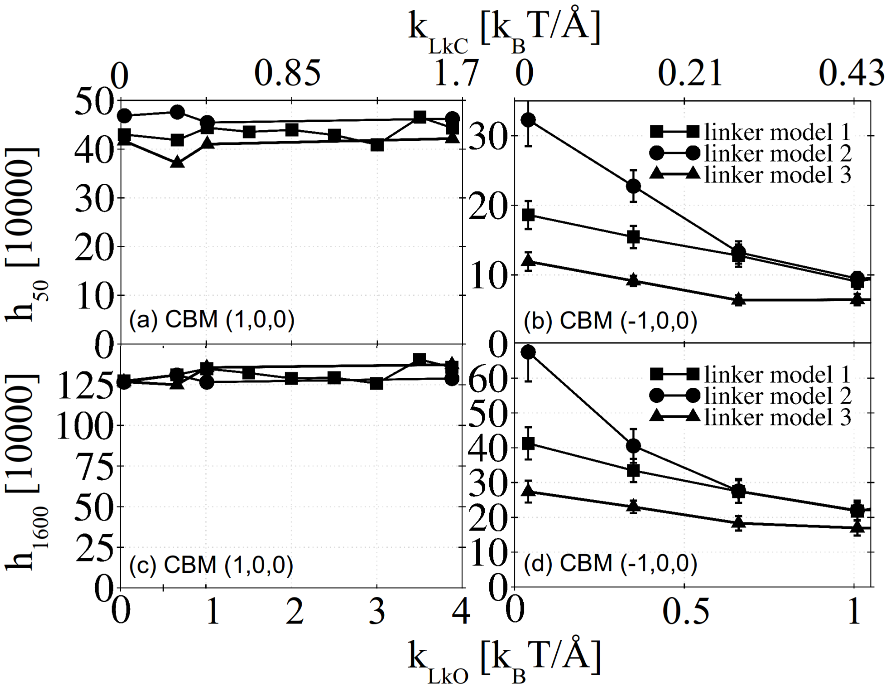


Figure 6.8: In the density map (e.g., Figure 4.14) the value of the sum of the 50 and 1600, respectively, most populated bins close to a given fiber face is denoted as h_{50} and h_{1600}, respectively. In sub figure a to d h_{50} resp. h_{1600} is plotted as a function of the linker stiffness. The BD trajectories are started above the hydrophobic fiber face (1, 0, 0) (left two figures a and c), the hydrophobic fiber face (-1, 0, 0) has the farthest distance from the starting position (right two figures c and d). In the figures b and d h_{50} and h_{1600}, respectively, decreases with increasing linker stiffness. This indicates that for a soft spring the CBM is more mobile and can visit more often the hotspot at (-1, 0, 0) (right two figures c and d).

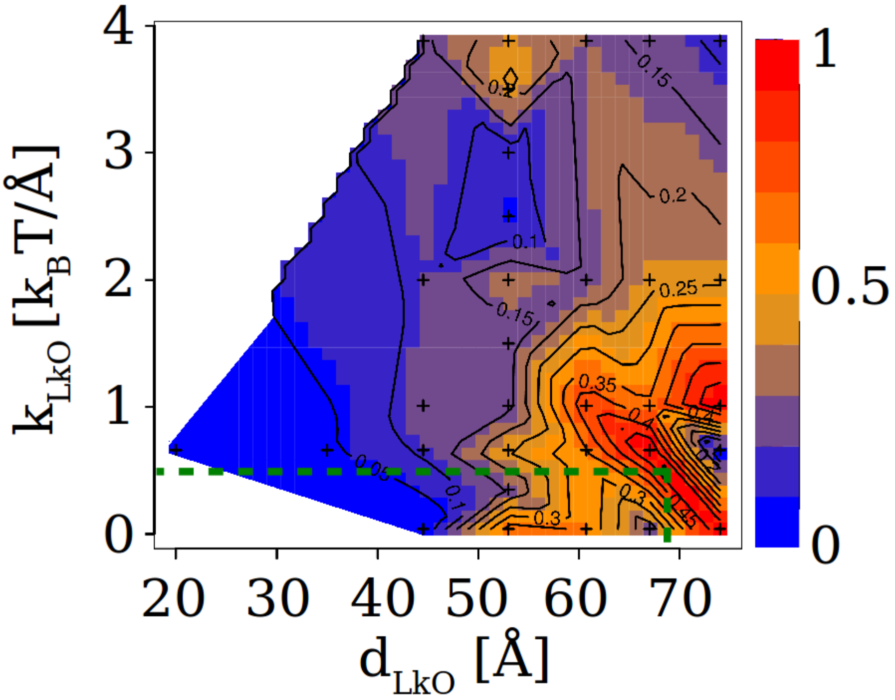


Figure 6.9: Visualization of $s(k_{LkO}, d_{LkO})$ as function of linker length d_{LkO} and stiffness k_{LkO} for the different BD simulation sets. The discrete values for d_{LkO} and k_{LkO} are marked with a plus. The intermediate values are interpolated with splines. $s(k_{LkO}, d_{LkO})$ is in the range of 0 to 1, where a value near 1 indicates an optimal value for the linker length and stiffness and 0 indicates a not optimal value. For the optimal value (marked in green) the CBM has an high mobility and the CD has an high probability to stay with the CBM on the same fiber face.

I am turned into a sort of machine for
observing facts and grinding out conclusions.

Charles Darwin 1809-1882

Chapter 7

Conclusion and Outlook

7.1 Conclusions

Cellulosic biomass has the potential to be a plentiful feedstock for the production of renewable biofuels. Cellulose fibers consist of linear chains of several hundred to over ten thousand linked sugar units. The basic idea is to break down the cellulose chain into the individual sugars using cellulase enzymes. This sugars can be fermented in the next step to bioethanol. For a successful shift from fossil fuels towards renewable biofuels from non-food cellulose waste, the productions costs have to be reduced by improving the enzymatic digestion. One of the promising enzymes for the enzymatic digestion of cellulose is Cel7A, due to its ability to bind and to disrupt the surface of crystalline cellulose. Cel7A consists of two domains, the carbohydrate-binding module (CBM) and a catalytic domain (CD), which are joined together by a linker peptide. Current industrial approaches require high enzymatic loads. Understanding the interaction of Cel7A with cellulose on an atomic level might help in the long term to experimentally design improved cellulase enzymes. Current experimental methods however do not have the required time and space resolution range, making it difficult to address this problem at an atomic level. Classical computer simulations on the other hand have the required time and space resolution to track each single atom during the Cel7A-cellulose interaction. But even on the most powerful supercomputers it is challenging to simulate trajectories which span the required time range of milliseconds. Using enhanced sampling methods like Markov state models and the combination of Brownian (BD) together with molecular dynamic (MD) simulations can help to cover the relevant time scales and lengths.
In this thesis various aspects of the Cel7A enzyme interaction with the cellulose fiber have been investigated using computer simulations, to address following central questions:

- *What is the role of the CBM? Which interactions forces are required to achieve the binding preferences of Cel7A to the fiber faces?*

- *What is the role of the linker peptide? What is the effect of the linker length* *and stiffness on the interaction of with cellulose fiber? What are the favored linker modes?*

To answer these questions, with the current level of knowledge and the available computational resources, an interdisciplinary combination of various methods from computer science, mathematics, and biophysics is required. Both questions have been discussed in detail in the Chapters 5 and 6. A summary of the major conclusions is presented here along with an outlook of possible further promising research directions.

7.1.1 Bridging the Gap between Theory and Experiments

One of the main aims of this thesis was to bridge the gap between experiments and theory for the Ce7A-cellulose system, by providing an atomic level a description of their interaction with computer simulations. This was achieved by simulating in total over 54,600 all-atom BD trajectories with a totaling $>$ 74 ms of combined simulation time. Additionally 101 all-atom MD trajectories with a total simulation time of $>$ 6.4 μs were simulated. In total this is longer then the previously reported simulations on this system, allowing me to give statistical more comprehensive analysis of the interaction of Cel7A with the cellulose Iβ.
The extensive MD simulations data presented provides means to test the parameters of the CHARMM27 force field for proteins [125], and the C35 force field for carbohydrates [129, 130]. The results presented in the thesis on hand show overall a good agreement with experiments.
The BD force field was extended by including the van der Walls (VDW) interaction between Cel7A and the cellulose fiber. The modification was verified, amongst others by calculating the density map of the CBM-fiber interaction, which shows CBM binding to similar fiber faces as previous experimental findings on cellulose Iα [40–42].

7.1.2 Overcoming the Time and Length Scale Problem for the Cel7A-Cellulose System

Cel7A-cellulose interactions take place on large time and length scale range, which is difficult to cover meaningful with only BD or only MD simulations. By combining in a multiscale approach BD with MD simulations I could cover the complete range from global (e.g. docking of the CBM to the fiber) to local interactions (e.g. side chain motions or hydrogen bond interactions).
By successfully developing a MSM framework for the Cel7A-cellulose system I could combine in a systematic fashion the statistical informations of independent simulation

trajectories. This allowed me to simulate instead of a single long trajectory, the required trajectories for the most part independently on a parallel supercomputer with nearly perfect scaling. This trick allowed me to reduce the total effective simulation time.

7.1.3 Extended Set of Coarse Grain Parameters

To use the limited computational resources more efficiently, instead of fine-grained but computational expensive MD simulations, computational cheaper coarse grain simulations can be performed. The approximations used in the coarse grain simulations will not significantly effect the results of interest. Using my extensive MD simulation data, I extended the set of available coarse grain parameters for the Cel7A-cellulose system, in particular the translational and rotation diffusion coefficients for the Cel7A and the fiber, and the linker length and stiffness. My parameters include the effect of explicit solvent molecules and most parameters are provided in dependence of the fiber surface hydrophobicity. This parameters will make more realistic coarse grain models of the Cel7A-fiber complex possible.

7.1.4 Cellulose Iβ

Native cellulose exists primarily in two different isoforms Iα and Iβ, for which the crystallographic unit cell differs. This might affect their interaction with the Cel7A enzyme. While most previous experiments [40–42, 186] have studied cellulose Iα, which is a predominant form in some algae and can be easily purified [39], the thesis in hand focuses on cellulose Iβ , which is the major constituent of higher plants and therefore the technologically more relevant cellulose isoform for the biofuel production. The calculated density map of the CBM-fiber interaction, as well as the MSM showed that the experimental results are transferable from Iα to Iβ. Moreover, I could quantify the CBM-fiber binding for each Iβ fiber face position, in particular the CBM accessibility to each fiber face, how “sticky” each fiber position is for the CBM, and the relative CBM orientations towards the fiber.

7.1.5 Role of Cel7A CBM

The first central theme of the thesis has been the Cel7A CBM. On several levels from the initial encounter of the free floating CBM with the cellulose fiber, diffusion on the fiber surface, and the binding to cellulose, a large body of previous work could be complemented and rationalized.

For the binding preference of the CBM to the hydrophobic fiber face it has been shown that both electrostatic (ES) and van der Waals (VDW) are required. At long distance the ES interaction is more important and aligns via a dipole-dipole interaction the CBM in an antiparallel manner relative to the fiber axis. In contrast, on short distances the VDW interaction plays a more dominant role by stabilizing the CBM-fiber binding, in addition the ES interaction contributes to the stabilization of the bind by the formation of hydrogen bonds between the CBM and the fiber. Moreover, an hindered diffusion of the CBM on all fiber faces was shown. The known hydrogen bond network between the CBM and the fiber was extended. Additional insights in the dynamic and role of the CBM residues Y5, Y13, Y31, and Y32 could be provided.

7.1.6 Role of Cel7A Linker Peptide

The second central theme of the thesis has been the Cel7A linker peptide. It was shown, that the linker equilibration length and stiffness are not intrinsic properties of the linker. They depend whether implicit or explicit solvent is used, whether the complete Cel7A or only the isolated linker is studied and the hydrophobicity of the fiber surface to which Cel7A is attached. The Cel7A linker has different states ranging from a compact to an extended conformation. The Cel7A packing in the compact conformation is hydrophobic as well as polar in nature. The linker length regulates the transition between the extended and compact linker states of the enzyme. The linker stiffness influences the CBM mobility, I furthermore quantified the effect of the linker stiffness on the CBM mobility. I determined the optimal linker length $\mathrm{d_{LkO}}$ and stiffness $\mathrm{k_{LkO}}$ for the CBM-fiber and CD-fiber interaction. In summary, the results show that the linker is not an accessory part of the Cel7A enzyme. The linker length and stiffness are not random properties, they rather have a significant effect on the Cel7A-fiber interaction. The new insights obtained here can assist in designing Cel7A with mutated linker that have improved hydrolysis properties, without having to alter the CBM or CD domain.

> We can only see a short distance ahead, but
> I can see plenty there that needs to be done.
>
> *Alan Turing 1912-1954*

7.2 Outlook

To obtain the results presented in this study a large set of simulation trajectories and a framework of analysis algorithms was developed. Together with the new insights presented in this thesis a new direction of research opens up to better understand on an atomic level the interaction of Cel7a with the cellulose fiber. A short summary of these follow-up studies is presented below.

7.2.1 Cel7A Basis-Conformation Decomposition

The results presented here showed that the Cel7A enzyme can adapt a large set of different states, ranging from compact to extended (Chapter 6). Biomolecules require solvent for their function. The Cel7A enzyme activity and its three dimensional structure depends on the pH level, which is effected by the surrounding solvent. The pH level can be adjusted to achieve optimal enzyme activity. The Cel7A tertiary structure can be studied using neutron scattering experiments. The direct output of such experiments is the scattering intensity profile $I(q)$ (Section 2.4). Such experiments are performed on an ensemble of Cel7A structures and they can last several hours, during which Cel7A can take on several different conformations from extended to compact. This two effects make the $I(q)$ data noisy and difficult to derive a representative three dimensional structure. Computer science can assist the experimentalist to interpret their intermediate scattering functions $I(q)$.
In the Cel7A MD trajectories several transition from the extended to the compact state have been observed. A set of n relevant Cel7A conformations can be extracted from this trajectories. For this conformations the neutron scattering experiment can be simulated on a supercomputer to get the corresponding $I(q)$ profile for each conformation. This simulated $I_{sim,i}(q)$ can serve as a basis set to decompose the experimental $I_{exp}(q)$ data (Figure 7.1). In linear algebra a vector can be represented as a linear combination of the basis vectors, similar $I_{exp}(q)$ can be described e.g. as a linear combination of $I_{sim,i}(q)$:

$$I_{exp}(q) = \sum_{i=1}^{n} p_i \cdot I_{sim,i}(q)\,, \tag{7.1}$$

with $\sum_{i=1}^{n} p_i = 1$ and $0 \leq p_i \leq 1$. An optimization algorithm like Levenberg-Marquardt [236] can be used to find the optimal probabilities p_i. This decomposition relates the unknown structural details during the experiment with the known three dimensional structure from the MD simulations. The obtained probabilities p_1 to p_n

give the contribution of each conformation in the experimental data. The tertiary structure of Cel7A bound to the cellulose fiber can be compared with the unbound Cel7A. The comparison can give new insights on how the structure of Cel7A is related to its function and how it effects the enzyme activity.

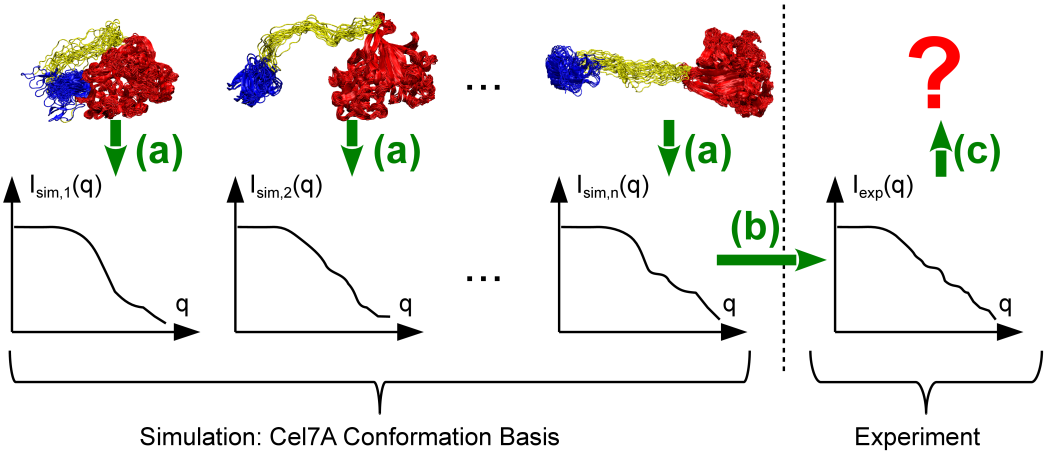


Figure 7.1: A three dimensional structure can be assigned to experimental $I_{exp}(q)$ data by using simulation data. In step (a) a basis set of representative Cel7A conformations are defined. For each conformation a simulated $I_{sim}(q)$ profile can be calculated. In step (b) and (c), this profiles can be combined to interpret $I_{exp}(q)$.

7.2.2 Error Minimization Sampling of Cel7A-Cellulose

Due to the limited computational resources, one of the major bottle necks in computer simulations is to obtain convergence of the derived results. For this thesis a MSM for the interaction of Cel7A enzyme with the cellulose fiber was developed utilizing BD and MD simulations. For large bimolecular systems like the Cel7A-cellulose complex this is a quite challenging task. In the MSM context observables can be calculated which describe the Cel7A-cellulose system. The amount of simulation data used is limited, therefore for each observable a statistical uncertainty can be calculated [167, 168]. The uncertainty can be reduced by increasing the simulation time. The MSM methodology allows to reconstruct the statistical informations of a single long trajectory from a large set of short trajectories, which can be simulated for the most part independently on a parallel supercomputer. Instead of extending a single trajectory which samples the Cel7A-cellulose conformations which cause the high uncertainty, for a given observable, shorter simulations can be systematically started from specific conformations (Figure 7.2). With such an iterative approach the statistical uncertainty for the values of interest can be reduced by using in total a smaller total simulation time compared to the case of an undirected single simulation trajectory.

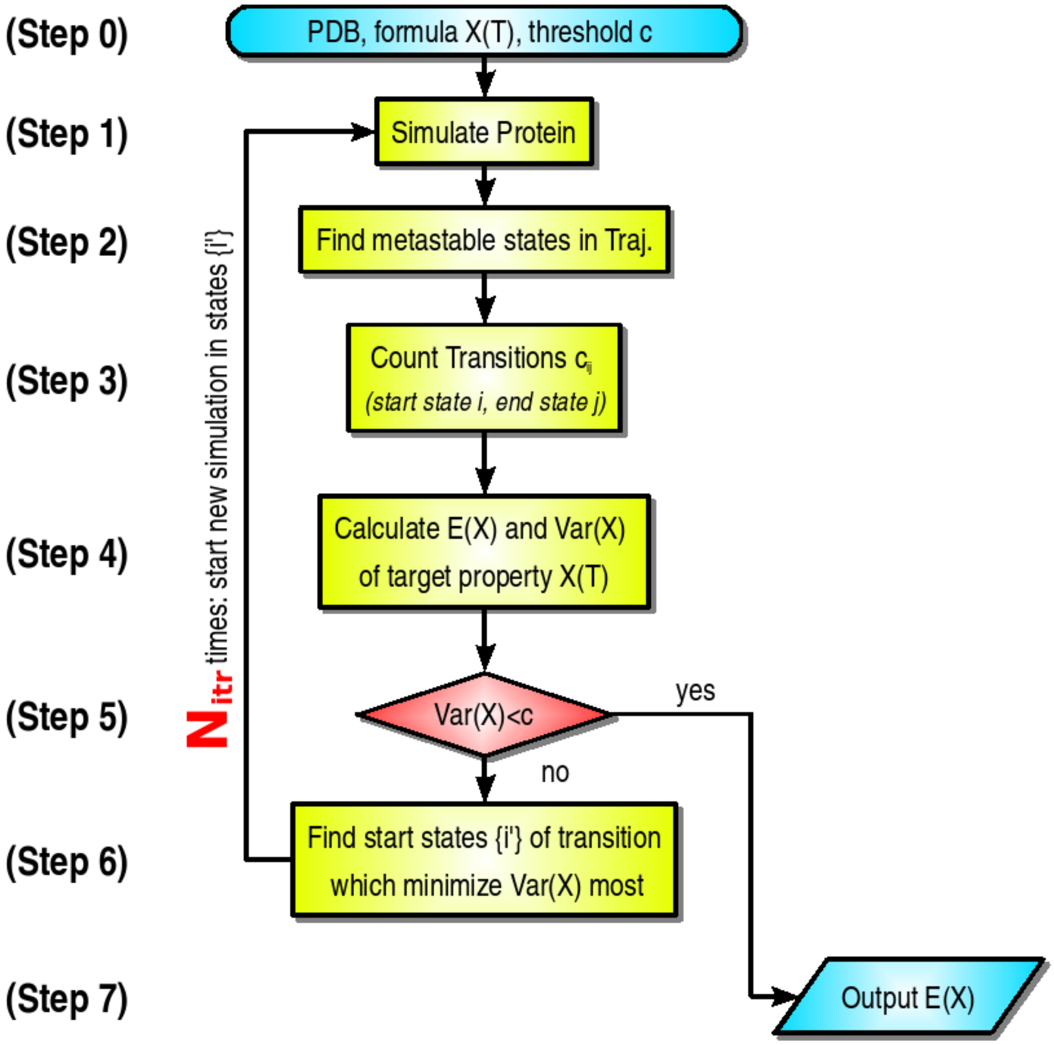

(Step 0)
(Step 1)
(Step 2)
(Step 3)
(Step 4)
(Step 5)
(Step 6)
(Step 7)
PDB, formula X(T), threshold c
Simulate Protein
Find metastable states in Traj.
Count Transitions c_{ij}
(start state i, end state j)
Calculate E(X) and Var(X)
of target property X(T)
Var(X)<c
yes
no
Find start states {i'} of transition
which minimize Var(X) most
Output E(X)
N_{itr} times: start new simulation in states {i'}

Figure 7.2: Using a MSM framework new simulations of the Cel7A-cellulose complex can be started iteratively to reduce the statistical uncertainty $Var(X)$ on the observable of interest X. The iteration process is terminated if the $Var(X)$ is smaller then a given threshold c.

7.2.3 Cel7A Motion on the Cellulose Fiber Surface

One of the essential steps during the enzymatic hydrolysis process is the motion of Cel7A via the CBM on the cellulose fiber surface. In this step, the enzyme scans the fiber surface for broken chain ends, which are in the next step threaded into the CD tunnel. The MD results reveal hindered diffusion of the CBM on all fiber surfaces. The diffusion motion is more restricted on the hydrophobic than on the hydrophilic fiber surface. To interpret the motion of the CBM a Brownian diffusion model was used, which is an oversimplification. More sophisticated approaches such as sub-diffusion and reaction-coupled diffusion models [206–208] can help to better understand the diffusion process of the Cel7A on the crystalline cellulose and give additional insights.

7.2.4 Nature of CBM-Cellulose Binding

Biomolecules require solvent to take on a three dimensional structure and to perform their function. The residues Y5, Y31, and Y32 constitute an hydrophobic patch at the bottom surface of the CBM. During the binding process of the CBM with the hydrophobic fiber surface, this two surfaces build an hydrophobic sandwich. The role of the solvent in the interface area is of special interest. Hydrogen bonding as well as hydrophobic interaction was observed between the CBM and the cellulose surface. The free energy of the binding

$$\triangle G = \triangle H - T\triangle S \tag{7.2}$$

can be decomposed into an enthalpic $\triangle H$ and an entropic part $T\triangle S$. To solvate an hydrophobic interface small water cavities have to be formed without interrupting hydrogen bonds, the free energy cost is therefore entropy dominated [237, 238]. The hydrogen bond interaction is rather enthalpy dominated [199, 239]. The questions remain, however what is the nature of the CBM-cellulose binding, is the binding more entropically or enthalpic driven? Further work is needed to answer this question.

It's the little details that are vital. Little things make big things happen.

John Wooden 1910-2010

Appendix A

Appendix

A.1 Abbreviations

- 36-chain: 36-chain cellulose fiber model
- BO: Born-Oppenheimer
- BD: Brownian dynamics
- CBM: carbohydrate binding module
- CD: catalytic domain
- CF: cellulose fiber
- ES: electrostatic interaction
- FT: Fourier transformation
- Lk: linker peptide
- MD: molecular dynamics
- MSM: Markov state model
- PDB: protein data bank
- PME: particle mesh Ewald
- PMF: potential of mean force
- SASA: solvent accessible surface area
- VDW: van der Waals interaction

A.2 Amino Acids

Amino acids are biologically important structural units that can make up proteins. The key chemical elements are carbon (C), hydrogen (H), oxygen (O), and nitrogen (N). They are build up from an *amine* ($-NH_2$), a *carboxylic axid* ($-COOHN$), and a side-chain molecule “R”. The generic structure is shown in Figure A.1. About 500 amino acids are known [240], but only 20 are encoded in the genetic code, they are called *standard amino acids*. The amino acids Y5, Y31, and Y31 of the Cel7A CBM domain constitute an hydrophobic patch (Section 2.2). The hydrophobicity [241] of the different amino acids is presented in Table A.1), Tyrosine (1-letter code “Y”) is with a value -1.3 rather hydrophobic.

Amino acid	3-letter	1-letter	Hydrophobicity [241]	R-group
Alanine	Ala	A	1.8	$-CH_3$
Arginine	Arg	R	-4.5	$CH_2CH_2CH_2NH-C(NH)NH_2$
Asparagine	Asn	N	-3.5	$-CH_2CONH_2$
Aspartic acid	Asp	D	-3.5	$-CH_2COOH$
Cysteine	Cys	C	2.5	$-CH_2SH$
Glutamic acid	Glu	E	-3.5	$-CH_2CH_2CONH_2$
Glutamine	Gln	Q	-3.5	$-CH_2CH_2COOH$
Glycine	Gly	G	-0.4	$-H$
Histidine	His	H	-3.2	$-CH_2(C_3H_3N_2)$
Isoleucine	Ile	I	4.5	$-CH(CH_3)CH_2CH_3$
Leucine	Leu	L	3.8	$-CH_2CH(CH_3)_2$
Lysine	Lys	K	-3.9	$-CH_2CH_2CH_2CH_2NH_2$
Methionine	Met	M	1.9	$-CH_2CH_2SCH_3$
Phenylalanine	Phe	F	2.8	$-CH_2(C_6H_5)$
Proline	Pro	P	-1.6	$-CH_2CH_2CH_2-$
Serine	Ser	S	-0.8	$-CH_2OH$
Threonine	Thr	T	-0.7	$-CH(OH)CH_3$
Tryptophan	Trp	W	-0.9	$-CH_2(C_8H_6N)$
Tyrosine	Tyr	Y	-1.3	$-CH_2(C_6H_4)OH$
Valine	Val	V	4.2	$-CH(CH_3)_2$

Table A.1: Overview of the standard amino acids.

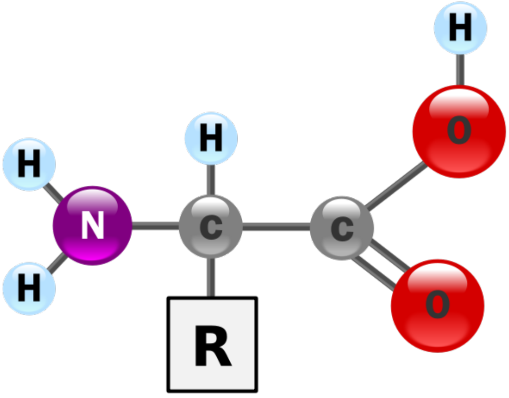


Figure A.1: Generic structure of an amino acid (adapted from [242]).

A.3 Glossary

This dissertation touches several different scientific fields. Each field, in particular biology, has its own set of vocabulary. The reader should not be discouraged by this fact. The focus of this glossary is to provide an understandable picture of a technical term.

- Alkenes: chemical compound with the general chemical structure C_nH_{2n}.
- Alpha chains: subunits of a protein.
- Ångström (Å): unit of length equal to 10^{10} m. Its symbol is the Swedish letter Å.
- Arabinose: special sugar molecule.
- Barrel (bbl): a oil barrel is 42 U.S. gallons or 158.9873 liter.
- Beta chains: subunits of a protein.
- Billion: the short scale naming system is used, a billion means 10^9.
- Biomass: raw plant material composed of glucose polymers (hemicellulose and cellulose) and lignin.
- Biomass recalcitrance: resistance of plant cell walls to decomposition.
- Brownian dynamics: a computer simulation method to simulate molecular interactions (Section 3.1).

- Carbohydrate: are a chemical compound with the chemical structure $C_m(H_2O)_n$.

- Carbohydrate-binding module (CBM): domain of the cellulase enzyme Cel7A (Section 2.2).

- Catalytic domain (CD): domain of the cellulase enzyme Cel7A (Section 2.2).

- Cel7A: special cellulase enzyme which is able to degrade cellulose efficiently (Section 2.2).

- Cellobiose unit: a chemical compound consisting of two linked sugar molecules (Section 2.1).

- Cellulase enzyme: enzymes that can degrade cellulose(Section 2.2).

- Cellulose fiber: special polymer of glucose molecules (Section 2.1).

- Deuterium: a heavy hydrogen, composed of a single neutron, a proton, and an electron.

- Dipole: see Section 2.5.

- Enzymatic digestion: breakdown of biomolecules (e.g., cellulose fiber) into simpler chemical compounds (Section 2.2).

- Enzyme: large biological molecules that are in many ways “protein machines” (Section 2.2).

- Ethanol: an alcohol with the chemical structure C_2H_6O.

- Galactose: special sugar molecule.

- Glucose: simple sugar with chemical structure $C_6H_{12}O_6$ (Section 2.1).

- Glycosidic bond: a chemical bond which joins a sugar molecules to another group (Section **??**).

- Glycosylation: enzymatic process that attaches sugars to proteins (Section 2.2).

- Heme group: chemical compound, Heme binds and carries oxygen in the red blood cells.

- Hemicellulose: special polymer of glucose molecules (Section 2.1).

- Hydrogen: chemical element with symbol H, consisting of a proton and an electron.

- Hydrophile: a molecule that is attracted to water.

- Hydrophobe: a molecule that is repelled from a mass of water.

- Lignin: is composed of a mixture of hydrophobic and aromatic molecules.

- Lignocellulose: a mixture of lignin, hemicellulose and cellulose is collectively called lignocellulose.

- Linker peptide: amino-acid chain of the cellulase enzyme Cel7A which is connected to the CD and CBM domain (Section 2.2).

- Mannose: special sugar molecule.

- Molecular dynamics: a computer simulation method to simulate molecular interactions (Section 3.2).

- Monosaccharides: simplest form of sugar, polymer containing only a single sugar.

- MSM: see Section 4.6.

- Oligosaccharide: polymer containing typically two to ten simple sugars.

- Peptide: short chains of amino acid.

- pH: is a measure of the hydrogen ion concentration. Pure water has a pH of approx. 7, acids have a pH less than 7, and basis have a pH greater than 7.

- Phenol ring: a chemical group with structure C_6H_5OH.

- Proteolysis: the breakdown of proteins into amino acids or smaller polypeptides.

- Rhamnose: special sugar molecule.

- Xylose: special sugar molecule.

A.4 References Quotes

- **List of Publications:**
  Anne Louise Germaine de Stal (22 April 1766 - 14 July 1817), commonly known as Madame de Stal. Quoted in "A Dictionary of Thoughts: Being a Cyclopedia of Laconic Quotations from the Best Authors, Both Ancient and Modern" (1891) edited by Tryon Edwards. p. 502.

- **Start Pages:**
  Albert Einstein (14 March 1879 - 18 April 1955). Attributed to Einstein by Frau Born. Paraphrased words as given in Ronald William Clark, "Einstein" (1984), p. 243.

  Blaise Pascal (16 June 1623 - 19 August 1662). Commonly attributed to Pascal. Provincial Letters: Letter XVI, 4 December, 1656 http://oregonstate.edu/instruct/phl302/texts/pascal/letters-c.html#LETTER%20XVI. The French "Je n'ai fait celle-ci plus longue que parce que je n'ai pas eu le loisir de la faire plus courte.". Literally: "I made this [letter] very long, because I did not have the leisure to make it shorter".

- **Chapter 1:**
  "Biofeuls: Challenges to the Transportation, Sale, and Use of Intermediate Ethanol Blends", United States Government Accountability Office, Report to Congressional Requesters, GAO-11-513 (Washington, D.C.: June 3, 2011), p. 2, 7, 11, http://gao.gov/assets/320/319297.pdf

- **Chapter 2:**
  Anne Campbell (9 April 1992 - 5 May 2005).

  Prof. J. Regalbuto, University of Illinois [13]

  Richard Feynman (11 May 1918 - 15 February 1988). Attributed to Feynman in the PBS TV show NOVA.

- **Chapter 3:**
  Richard Feynman (11 May 1918 - 15 February 1988). The Feynman Lectures on Physics.

  John von Neumann (28 December 1903 - 18 February 1957).

  Paul A. M. Dirac (8 August 1902 - 20 October 1984).

- **Chapter 4:**
  Prof. Dr. Jörg Hüfner, University Heidelberg, Institute for Theoretical Physics. He gave this advice during one of his theoretical physics lectures.

  Master Zhuang (369 BC - 286 BC)

  Albert Einstein (14 March 1879 - 18 April 1955). Attributed to Einstein by Readers Digest in July 1977

  Albert Einstein (14 March 1879 - 18 April 1955).

  Johann Wolfgang von Goethe (28 August 1749 22 March 1832). Letter to Johann Christian Lobe from July 1820.

  Jack J. Dongarra (born 18 July 1950). Professor of Computer Science at the University of Tennessee. Dongarra holds the Turing Fellowship at the University of Manchester. He is one of the compilers of the ”TOP500” project, which ranks the 500 most powerful computer systems in the world.

- **Chapter 5:**
  Aristotle (384 BC - 322 BC). Short form of a quote from “Metaphysics VII”.

- **Chapter 6:**
  Robert Hooke (28 July 1635 - 3 March 1703). Short form of a quote from “Micrographia 1665”.

- **Chapter 7:**
  Charles Robert Darwin (12 February 1809 - 19 April 1882). As quoted in Adrian J. Desmond and James Richard Moore, ”Darwin” (1994), p. 644.

- **Appendix:**
  John Robert Wooden (14 October 1910 - 4 June 2010).

  Alan Turing (23 June 1912 - 7 June 1954). Published in the article ”Computing machinery and intelligence” (Mind, vol. 59, 1950). This paper describes what later was to be known as the ”Turing Test”, http://www.loebner.net/Prizef/TuringArticle.html.

References

[1] Martin Robbins. Policy: Fuelling politics. *Nature*, 474(7352):S22–S24, 2011.

[2] José Goldemberg and Suani Teixeira Coelho. Renewable energytraditional biomass vs. modern biomass. *Energy Policy*, 32(6):711–714, 2004.

[3] Duncan Graham-Rowe. Agriculture: Beyond food versus fuel. *Nature*, 474(7352):S6–S8, 2011.

[4] Antoinette C. Sullivan. Cellulose: the structure slowly unravels. *Cellulose*, 4:173–207, 1997.

[5] J.J. Meister. *Polymer Modification: Principles, Techniques, and Applications.* CRC, 2000.

[6] P. Falkowski, R. J. Scholes, E. Boyle, J. Canadell, D. Canfield, J. Elser, N. Gruber, K. Hibbard, P. Hgberg, S. Linder, F. T. Mackenzie, B. Moore, T. Pedersen, Y. Rosenthal, S. Seitzinger, V. Smetacek, and W. Steffen. The global carbon cycle: a test of our knowledge of earth as a system. *Science*, 290(5490):291–296, Oct 2000.

[7] MT Holtzapple. Cellulose in, macrae r, robinson rk, saddler mj, editors. encyclopedia of food science food technology and nutrition. *London: Academic Press*, 16:758–767, 1993.

[8] M. Jarvis. Cellulose stacks up. *Nature*, 426(6967):611–612, 2003.

[9] R.M. Brown. *Cellulose structure and biosynthesis: What is in store for the 21st century?*, volume 42. [New York, NY]: Wiley,[c1986-, 2004.

[10] Marcos Silveira Buckeridge. *Routes to cellulosic ethanol.* Springer Science+ Business Media, 2011.

[11] Peter Fairley. Introduction: Next generation biofuels. *Nature*, 474(7352):S2–S5, 2011.

[12] Katharine Sanderson. Lignocellulose: A chewy problem. *Nature*, 474(7352):S12–S14, 2011.

[13] N. Savage. Fuel options: The ideal biofuel. *Nature*, 474(7352):S9–S11, 2011.

[14] Neil Savage. Algae: The scum solution. *Nature*, 474(7352):S15–S16, 2011.

[15] L.R. Lynd and J. Woods. Perspective: A new hope for africa. *Nature*, 474(7352):S20–S21, 2011.

[16] Marcia Moraes. Perspective: lessons from brazil. *Nature*, 474(7352):S25–S25, 2011.

[17] J. Martin. Perspective: Don't foul the water. *Nature*, 474(7352):S17–S17, 2011.

[18] José Goldemberg et al. The brazilian biofuels industry. *Biotechnology for Biofuels*, 1(6):1–7, 2008.

[19] José Goldemberg. The brazilian experience with biofuels (innovations case narrative). *Innovations: Technology, Governance, Globalization*, 4(4):91–107, 2009.

[20] Jose Goldemberg. The role of biomass in the worlds energy system. In Marcos Silveira Buckeridge and Gustavo H Goldman, editors, *Routes to Cellulosic Ethanol*, pages 3–14. Springer New York, 2011.

[21] Christina L Ting, Dmitrii E Makarov, and Zhen-Gang Wang. A kinetic model for the enzymatic action of cellulase. *The journal of physical chemistry B*, 113(14):4970–4977, 2009.

[22] European Commission et al. Biofuels in the european union: A vision for 2030 and beyond. *Final report of the Biofuels Research Advisory Council, EUR*, 22066, 2006.

[23] M.E. Himmel, S.Y. Ding, D.K. Johnson, W.S. Adney, M.R. Nimlos, J.W. Brady, and T.D. Foust. Biomass recalcitrance: engineering plants and enzymes for biofuels production. *Science*, 315(5813):804, 2007.

[24] José Goldemberg and Patricia Guardabassi. Are biofuels a feasible option? *Energy Policy*, 37(1):10–14, 2009.

[25] T.D. Foust, A. Aden, A. Dutta, and S. Phillips. An economic and environmental comparison of a biochemical and a thermochemical lignocellulosic ethanol conversion processes. *Cellulose*, 16(4):547–565, 2009.

[26] P Pepiot, CJ Dibble, and TD Foust. Computational fluid dynamics modeling of biomass gasification and pyrolysis. *Computational modeling in lignocellulosic biofuel production, American Chemical Society, Golden*, pages 273–298, 2010.

[27] Tiziano Gomiero, Maurizio Paoletti, and David Pimentel. Biofuels: Efficiency, ethics, and limits to human appropriation of ecosystem services. *Journal of Agricultural and Environmental Ethics*, 23:403–434, 2010. 10.1007/s10806-009-9218-x.

[28] George W. Huber and Bruce E. Dale. Erde 3.0: Grasolin an der zapfsule. *Spektrum der Wissenschaft*, Dezember:88–94, 2009.

[29] M. R. Schmer, K. P. Vogel, R. B. Mitchell, and R. K. Perrin. Net energy of cellulosic ethanol from switchgrass. *Proceedings of the National Academy of Sciences*, 105(2):464–469, 2008.

[30] Chris Somerville, Stefan Bauer, Ginger Brininstool, Michelle Facette, Thorsten Hamann, Jennifer Milne, Erin Osborne, Alex Paredez, Staffan Persson, Ted Raab, Sonja Vorwerk, and Heather Youngs. Toward a Systems Approach to Understanding Plant Cell Walls. *Science*, 306(5705):2206–2211, 2004.

[31] Meijuan Zeng. *Characterization of cell wall deconstruction induced by aqueous pretreatment and enzyme hydrolysis.* ProQuest, 2007.

[32] Chen Zhang and Se-Kwon Kim. Research and application of marine microbial enzymes: status and prospects. *Marine drugs*, 8(6):1920–1934, 2010.

[33] Shulin Chen, Xiaoyu Zhang, Deepak Singh, Hongbo Yu, and Xuewei Yang. Biological pretreatment of lignocellulosics: potential, progress and challenges. *Biofuels*, 1(1):177–199, 2010.

[34] Á. Tímár-Balázsy and D. Eastop. *Chemical principles of textile conservation.* Butterworth-Heinemann, 1998.

[35] John M Ziman. *Principles of the Theory of Solids.* Cambridge University Press, 1979.

[36] R.H. Atalla and D.L. Vanderhart. Native cellulose: a composite of two distinct crystalline forms. *Science*, 223:283–285, 1984.

[37] J. Sugiyama, R. Vuong, and H. Chanzy. Electron diffraction study on the two crystalline phases occurring in native cellulose from an algal cell wall. *Macromolecules*, 24(14):4168–4175, 1991.

[38] Yoshiharu Nishiyama, Paul Langan, and Henri Chanzy. Crystal structure and hydrogen-bonding system in cellulose ibeta from synchrotron x-ray and neutron fiber diffraction. *J Am Chem Soc*, 124(31):9074–9082, Aug 2002.

[39] Yoshiharu Nishiyama, Junji Sugiyama, Henri Chanzy, and Paul Langan. Crystal structure and hydrogen bonding system in cellulose i(alpha) from synchrotron x-ray and neutron fiber diffraction. *J Am Chem Soc*, 125(47):14300–14306, Nov 2003.

[40] Janne Lehtiö, Junji Sugiyama, Malin Gustavsson, Linda Fransson, Markus Linder, and Tuula T. Teeri. The binding specificity and affinity determinants of family 1 and family 3 cellulose binding modules. *Proceedings of the National Academy of Sciences of the United States of America*, 100(2):484–489, 2003.

[41] D.J. Dagel, Y.S. Liu, L. Zhong, Y. Luo, M.E. Himmel, Q. Xu, Y. Zeng, S.Y. Ding, and S. Smith. In Situ Imaging of Single Carbohydrate-Binding Modules on Cellulose Microfibrils. *The Journal of Physical Chemistry B*, 115:635–641, 2010.

[42] Y.S. Liu, J.O. Baker, Y. Zeng, M.E. Himmel, T. Haas, and S.Y. Ding. Cellobiohydrolase hydrolyzes crystalline cellulose on hydrophobic faces. *Journal of Biological Chemistry*, 286(13):11195, 2011.

[43] L. Zhong, J.F. Matthews, M.F. Crowley, T. Rignall, C. Talón, J.M. Cleary, R.C. Walker, G. Chukkapalli, C. McCabe, M.R. Nimlos, et al. Interactions of the complete cellobiohydrolase I from Trichodera reesei with microcrystalline cellulose Iβ. *Cellulose*, 15(2):261–273, 2008. Cellulose potential energy is stored in its C-H and C-C bonds.

[44] University of windsor, computational chemistry.

[45] Charlotte Schubert. Can biofuels finally take center stage? *Nat Biotechnol*, 24(7):777–784, Jul 2006.

[46] W. Haynes. *Cellulose, the chemical that grows.* Doubleday, 1953.

[47] Shi-You Ding and Michael E Himmel. Anatomy and ultrastructure of maize cell walls: an example of energy plants. *Biomass recalcitrance: deconstructing the plant cell wall for bioenergy*, pages 38–60, 2008.

[48] Tuula T. Teeri. Crystalline cellulose degradation: new insight into the function of cellobiohydrolases. *Trends in Biotechnology*, 15(5):160 – 167, 1997. Reference for exo- and endcellulase.

[49] B.K. Barr, Y.L. Hsieh, B. Ganem, and D.B. Wilson. Identification of two functionally different classes of exocellulases. *Biochemistry*, 35(2):586–592, 1996.

[50] M. J. Harrison, A. S. Nouwens, D. R. Jardine, N. E. Zachara, A. A. Gooley, H. Nevalainen, and N. H. Packer. Modified glycosylation of cellobiohydrolase i from a high cellulase-producing mutant strain of trichoderma reesei. *Eur J Biochem*, 256(1):119–127, Aug 1998.

[51] Xiongce Zhao, Tauna R Rignall, Clare McCabe, William S Adney, and Michael E Himmel. Molecular simulation evidence for processive motion of¡ i¿ trichoderma reesei¡/i¿ cel7a during cellulose depolymerization. *Chemical Physics Letters*, 460(1):284–288, 2008.

[52] M.E. Himmel. *Biomass recalcitrance: deconstructing the plant cell wall for bioenergy.* Blackwell Pub., 2008.

[53] J. Kraulis, G. M. Clore, M. Nilges, T. A. Jones, G. Pettersson, J. Knowles, and A. M. Gronenborn. Determination of the three-dimensional solution structure of the c-terminal domain of cellobiohydrolase i from trichoderma reesei. a study using nuclear magnetic resonance and hybrid distance geometry-dynamical simulated annealing. *Biochemistry*, 28(18):7241–7257, Sep 1989.

[54] A. M. Hoffrn, T. T. Teeri, and O. Teleman. Molecular dynamics simulation of fungal cellulose-binding domains: differences in molecular rigidity but a preserved cellulose binding surface. *Protein Eng*, 8(5):443–450, May 1995.

[55] T. Reinikainen, L. Ruohonen, T. Nevanen, L. Laaksonen, P. Kraulis, T. A. Jones, J. K. Knowles, and T. T. Teeri. Investigation of the function of mutated cellulose-binding domains of trichoderma reesei cellobiohydrolase i. *Proteins*, 14(4):475–482, Dec 1992.

[56] M. Srisodsuk, J. Lehtiö, M. Linder, E. Margolles-Clark, T. Reinikainen, and T. T. Teeri. Trichoderma reesei cellobiohydrolase i with an endoglucanase cellulose-binding domain: action on bacterial microcrystalline cellulose. *J Biotechnol*, 57(1-3):49–57, Sep 1997.

[57] T. Nagy, P. Simpson, M.P. Williamson, G.P. Hazlewood, H.J. Gilbert, and L. Orosz. All three surface tryptophans in Type IIa cellulose binding domains play a pivotal role in binding both soluble and insoluble ligands. *FEBS letters*, 429(3):312–316, 1998.

[58] A.B. Boraston, D.N. Bolam, H.J. Gilbert, and G.J. Davies. Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochemical Journal*, 382(Pt 3):769, 2004.

[59] D. Guillén, S. Sánchez, and R. Rodríguez-Sanoja. Carbohydrate-binding domains: multiplicity of biological roles. *Applied microbiology and biotechnology*, 85(5):1241–1249, 2010.

[60] C. Hervé, A. Rogowski, A.W. Blake, S.E. Marcus, H.J. Gilbert, and J.P. Knox. Carbohydrate-binding modules promote the enzymatic deconstruction of intact plant cell walls by targeting and proximity effects. *Proceedings of the National Academy of Sciences*, 107(34):15293–15298, 2010.

[61] J. Tormo, R. Lamed, A.J. Chirino, E. Morag, E.A. Bayer, Y. Shoham, and T.A. Steitz. Crystal structure of a bacterial family-iii cellulose-binding domain: a general mechanism for attachment to cellulose. *The EMBO Journal*, 15(21):5739, 1996.

[62] M.R. Nimlos, J.F. Matthews, M.F. Crowley, R.C. Walker, G. Chukkapalli, J.W. Brady, W.S. Adney, J.M. Cleary, L. Zhong, and M.E. Himmel. Molecular modeling suggests induced fit of Family I carbohydrate-binding modules with a broken-chain cellulose surface. *Protein Engineering Design and Selection*, 20(4):179, 2007.

[63] L. Bu, G.T. Beckham, M.R. Shirts, M.R. Nimlos, W.S. Adney, M.E. Himmel, and M.F. Crowley. Probing carbohydrate product expulsion from a processive cellulase with multiple absolute binding free energy methods. *Journal of Biological Chemistry*, -:–, 2011.

[64] Carl S Rye and Stephen G Withers. Glycosidase mechanisms. *Current opinion in chemical biology*, 4(5):573–580, 2000.

[65] Klaus Klarskov, Kathleen Piens, Jerry Ståhlberg, Peter B Høj, JV Beeumen, and Marc Claeyssens. Cellobiohydrolase i from trichoderma reesei: identification of an active-site nucleophile and additional information on sequence including the glycosylation pattern of the core protein. *Carbohydrate research*, 304(2):143–154, 1997.

[66] J. P. Hui, P. Lanthier, T. C. White, S. G. McHugh, M. Yaguchi, R. Roy, and P. Thibault. Characterization of cellobiohydrolase i (cel7a) glycoforms from extracts of trichoderma reesei using capillary isoelectric focusing and electrospray mass spectrometry. *J Chromatogr B Biomed Sci Appl*, 752(2):349–368, Mar 2001. Ref paper for sequence of linker.

[67] W.S. Adney, T. Jeoh, G.T. Beckham, Y.C. Chou, J.O. Baker, W. Michener, R. Brunecky, and M.E. Himmel. Probing the role of n-linked glycans in the stability and activity of fungal cellobiohydrolases by mutational analysis. *Cellulose*, 16(4):699–709, 2009.

[68] T. Jeoh, W. Michener, M.E. Himmel, S.R. Decker, and W.S. Adney. Implications of cellobiohydrolase glycosylation for use in biomass conversion. *Biotechnology for Biofuels*, 1(1):10, 2008.

[69] I. Stals, K. Sandra, S. Geysens, R. Contreras, J. Van Beeumen, and M. Claeyssens. Factors influencing glycosylation of trichoderma reesei cellulases. i: Postsecretorial changes of the o-and n-glycosylation pattern of cel7a. *Glycobiology*, 14(8):713, 2004.

[70] I. Stals, K. Sandra, B. Devreese, J. Van Beeumen, and M. Claeyssens. Factors influencing glycosylation of trichoderma reesei cellulases. ii: N-glycosylation of cel7a core protein isolated from different strains. *Glycobiology*, 14(8):725, 2004.

[71] S. Godbole, S.R. Decker, R.A. Nieves, W.S. Adney, T.B. Vinzant, J.O. Baker, S.R. Thomas, and M.E. Himmel. Cloning and expression of trichoderma reesei cellobiohydrolase i in pichia pastoris. *Biotechnology progress*, 15(5):828–833, 1999.

[72] S.P. Voutilainen, P.G. Murray, M.G. Tuohy, and A. Koivula. Expression of talaromyces emersonii cellobiohydrolase cel7a in saccharomyces cerevisiae and rational mutagenesis to improve its thermostability and activity. *Protein Engineering Design and Selection*, 23(2):69, 2010.

[73] H. Nevalainen, M. Harrison, D. Jardine, N. Zachara, M. Paloheimo, P. Suominen, A. Gooley, and N. Packer. Glycosylation of cellobiohydrolase i from trichoderma reesei. *Carbohydrates from Trichoderma reesei and other microorganisms: structures, biochemistry, genetics and applications*, 219:335–344, 1998.

[74] M. Srisodsuk, T. Reinikainen, M. Penttil, and T. T. Teeri. Role of the interdomain linker peptide of trichoderma reesei cellobiohydrolase i in its interaction with crystalline cellulose. *J Biol Chem*, 268(28):20756–20761, Oct 1993.

[75] M.L. Langsford, N.R. Gilkes, B. Singh, B. Moser, R.C. Miller jr, R.A.J. Warren, and D.G. Kilburn. Glycosylation of bacterial cellulases prevents proteolytic cleavage between functional domains. *FEBS Letters*, 225(1-2):163–167, 1987.

[76] A.J. Clarke. *Biodegradation of cellulose: enzymology and biotechnology.* CRC, 1996.

[77] X. Zhao, T.R. Rignall, C. McCabe, W.S. Adney, and M.E. Himmel. Molecular simulation evidence for processive motion of Trichoderma reesei Cel7A during cellulose depolymerization. *Chemical Physics Letters*, 460(1-3):284–288, 2008.

[78] I. Von Ossowski, J.T. Eaton, M. Czjzek, S.J. Perkins, T.P. Frandsen, M. Sch "ulein, P. Panine, B. Henrissat, and V. Receveur-Bréchot. Protein disorder: conformational distribution of the flexible linker in a chimeric double cellulase. *Biophysical journal*, 88(4):2823–2832, 2005.

[79] Vronique Receveur, Mirjam Czjzek, Martin Schlein, Pierre Panine, and Bernard Henrissat. Dimension, shape, and conformational flexibility of a two domain fungal cellulase in solution probed by small angle x-ray scattering. *J Biol Chem*, 277(43):40887–40892, Oct 2002.

[80] Badal C Saha and Jonathan Woodward. *Fuels and chemicals from biomass.* American Chemical Society, 1997.

[81] M. Bée. *Quasielastic neutron scattering: Principles and applications in solid state chemistry, biology and materials science.* Institute of Physics Publishing, isbn 0-8527-4371-8 edition, 1988.

[82] J.C. Smith. Protein dynamics: comparison of simulation with inelastic neutron experiments. *Quarterly Reviews of Biophysics*, 24(3):227–291, 1991.

[83] D. Bicout U. Lehnert M. Tehei M. Weik Gabel, F. and G. Zaccai. Protein dynamics studied by neutron scattering. *Quarterly Reviews of Biophysics*, 35(4):327–367, 2002.

[84] L. van Hove. Correlations in space and time and born approximation scattering in szstems of interacting particles. *Physcial Review*, 95(1), 1954.

[85] Roger Pynn. The mathematical foundations of neutron scattering, los alamos science summer 1990.

[86] Kei Moritsugu and Jeremy C. Smith. Langevin model of the temperature and hydration dependence of protein vibrational dynamics. *J. Phys. Chem. B*, 109:12182–12194, 2005.

[87] D. Bicout U. Lehnert M. Tehei M. Weik Gabel, F. and G. Zaccai. Protein dynamics studied by neutron scattering. *Quarterly Reviews of Biophysics*, 35(4):327–367, 2002. neutronScatteringReviewZaccai.

[88] Emal Aelozai. Protein dynamics in d2o and h2o: Solvation and isotope effects. *Diploma thesis, Faculty of Physics, Heidelberg University, Germany*, 2006.

[89] Roger Pynn. Neutron scattering - a primer. *Los Alamos Science*, 19, 1990.

[90] Hoover WG. Cononical dynamics: equilibrium phase space distributions. *Phys. Rev. A: At. Mol. Opt. Phys*, 31:1695–1697, 1985.

[91] Sing T. Bow. *Pattern Recognition and Image Preprocessing.* Signal Processing and Communication 14, isbn 0-8247-0659-5 edition, 2002. ISBN 0-8247-0659-5.

[92] Konstantinos Koutroumbas Sergios Theodoridis. *Pattern Recognition.* Academic Press, 3 edition, 1999. ISBN 0-1236-9531-7.

[93] Mathworld: Fourier transform.

[94] Stéphane Mallat. *A Wavlet tour of signal precessing.* Academic Press, 2 edition, 1999. ISBN 0-1246-6606-X.

[95] Bernd Jähne. *Digital Image Processing.* Springer, 2002. ISBN 3-540-67754-2.

[96] M.P. Allen and D.J. Tildesley. *Computer simulation of liquids.* Clarendon Press, 1989.

[97] Andrew R. Leach. *Molecular modelling: principles and applications.* Pearson College Division, 2001.

[98] Daan Frenkel and Berend Smit. *Understanding molecular simulation: from algorithms to applications.* Academic press, 2001.

[99] Tamar Schlick. *Molecular modeling and simulation: an interdisciplinary guide*, volume 21. Springer, 2010.

[100] Frank Jensen. *Introduction to computational chemistry.* Wiley, 2007.

[101] Dennis C Rapaport. *The art of molecular dynamics simulation.* Cambridge university press, 2004.

[102] A.H. Elcock, R.R. Gabdoulline, R.C. Wade, and J.A. McCammon. Computer simulation of protein-protein association kinetics: acetylcholinesterase-fasciculin1. *Journal of molecular biology*, 291(1):149–162, 1999.

[103] R.R. Gabdoulline and R.C. Wade. Biomolecular diffusional association. *Current opinion in structural biology*, 12(2):204–213, 2002.

[104] Razif R Gabdoulline and Rebecca C Wade. On the contributions of diffusion and thermal activation to electron transfer between phormidium laminosum plastocyanin and cytochrome f: Brownian dynamics simulations with explicit modeling of nonpolar desolvation interactions and electron transfer events. *Journal of the American Chemical Society*, 131(26):9230–9238, 2009.

[105] Alexander Spaar and Volkhard Helms. Free energy landscape of protein-protein encounter resulting from brownian dynamics simulations of barnase: barstar. *Journal of Chemical Theory and Computation*, 1(4):723–736, 2005.

[106] RR Gabdoulline and RC Wade. Effective charges for macromolecules in solvent. *J. Phys. Chem*, 100(9):3868–3878, 1996.

[107] J. Warwicker and HC Watson. Calculation of the electric potential in the active site cleft due to [alpha]-helix dipoles. *Journal of Molecular Biology*, 157(4):671–679, 1982.

[108] I. Klapper, R. Hagstrom, R. Fine, K. Sharp, and B. Honig. Focusing of electric fields in the active site of Cu-Zn superoxide dismutase: effects of ionic strength and amino-acid modification. *Proteins: Structure, Function, and Bioinformatics*, 1(1):47–59, 1986.

[109] ME Davis and J Andrew McCammon. Calculating electrostatic forces from grid-calculated potentials. *Journal of Computational Chemistry*, 11(3):401–409, 1990.

[110] Jeffry D. Madura, James M. Briggs, Rebecca C. Wade, Malcolm E. Davis, Brock A. Luty, Andrew Ilin, Jan Antosiewicz, Michael K. Gilson, Babak Bagheri, L. Ridgway Scott, and J. Andrew McCammon. Electrostatics and diffusion of molecules in solution: simulations with the university of houston brownian dynamics program. *Computer Physics Communications*, 91(1-3):57 – 95, 1995.

[111] N. A. Baker, D. Sept, S. Joseph, M. J. Holst, and J. A. McCammon. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc Natl Acad Sci U S A*, 98(18):10037–10041, Aug 2001.

[112] G. Akerlof. Dielectric constants of some organic solvent-water mixtures at various temperatures. *Journal of the American Chemical Society*, 54(11):4125–4139, 1932.

[113] M. Uematsu and EU Franck. *Static dielectric constant of water and steam.* National Standard Reference Data System, 1980.

[114] D.E. Kane. The relationship between the dielectric constant and water-vapor accessibility of cellulose. *Journal of Polymer Science*, 18(89):405–410, 2003.

[115] Clipper Controls. Dielectric constant reference guide, 2010.

[116] T.J. Dolinsky, J.E. Nielsen, J.A. McCammon, and N.A. Baker. PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic acids research*, 32(Web Server Issue):W665, 2004.

[117] Barry D. Olafson David J. States S. Swaminathan B. R. Brooks, Robert E. Bruccoleri and Martin Karplus. Charmm: A programm for macromolecular energy, minimization and dynamics calculations. *Journal of Computational Chemistry*, 4(2):187–217, 1983.

[118] R. Palma, P. Zuccato, M.E. Himmel, G. Liang, and J.W. Brady. Molecular mechanics studies of cellulases. In *ACS Symposium Series*, volume 769, pages 112–130. ACS Publications, 2001.

[119] M. Kuttel, JW Brady, and K.J. Naidoo. Carbohydrate solution simulations: producing a force field with experimentally consistent primary alcohol rotational frequencies and populations. *Journal of computational chemistry*, 23(13):1236–1243, 2002.

[120] M. Karttunen, I. Vattulainen, and A. Lukkarinen. *Novel methods in soft matter simulations.* Springer Verlag, 2004.

[121] M. Griebel, S. Knapek, and G. Zumbusch. *Numerical simulation in molecular dynamics.* Springer Berlin, 2007.

[122] D.L. Ermak and JA McCammon. Brownian dynamics with hydrodynamic interactions. *The Journal of Chemical Physics*, 69:1352, 1978.

[123] M Born and R Oppenheimer. On the quantum theory of molecules. *Annalen der Physik*, 84(20):457–484, 1927.

[124] H Köuppel, W Domcke, and LS Cederbaum. Multimode molecular dynamics beyond the born-oppenheimer approximation. *Advances in Chemical Physics, Volume 57*, pages 59–246, 2007.

[125] A. D. MacKerell, N. Banavali, and N. Foloppe. Development and current status of the charmm force field for nucleic acids. *Biopolymers*, 56(4):257–265, 2000.

[126] Lars Meinhold. Crystalline protein dynamics: A simulation analysis of stahphylococcal nuclease. *PhD Thesis, Faculty of Physics, Heidelberg University, Germany*, 2005.

[127] Loup Verlet. Computer" experiments" on classical fluids. i. thermodynamical properties of lennard-jones molecules. *Physical review*, 159(1):98, 1967.

[128] J.C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R.D. Skeel, L. Kale, and K. Schulten. Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry*, 26(16):1781, 2005.

[129] Olgun Guvench, Shannon N Greene, Ganesh Kamath, John W Brady, Richard M Venable, Richard W Pastor, and Alexander D Mackerell. Additive empirical force field for hexopyranose monosaccharides. *J Comput Chem*, 29(15):2543–2564, Nov 2008.

[130] Olgun Guvench, Elizabeth R Hatcher, Richard M Venable, Richard W Pastor, and Alexander D Mackerell. Charmm additive all-atom force field for glycosidic linkages between hexopyranoses. *J Chem Theory Comput*, 5(9):2353–2370, Aug 2009.

[131] W.L. Jorgensen, J. Chandrasekhar, J.D. Madura, R.W. Impey, and M.L. Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics*, 79:926, 1983.

[132] T. Darden, D. York, and L. Pedersen. Particle mesh Ewald: An N log (N) method for Ewald sums in large systems. *The Journal of Chemical Physics*, 98:10089, 1993.

[133] S.E. Feller, Y. Zhang, R.W. Pastor, and B.R. Brooks. Constant pressure molecular dynamics simulation: the Langevin piston method. *The Journal of chemical physics*, 103:4613, 1995.

[134] Ninad Prabhu and Kim Sharp. Protein-solvent interactions. *Chemical reviews*, 106(5):1616, 2006.

[135] H.M. Senn and W. Thiel. Qm/mm studies of enzymes. *Current opinion in chemical biology*, 11(2):182–187, 2007.

[136] Hans Senn and Walter Thiel. Qm/mm methods for biological systems. *Atomistic approaches in modern biology*, 268:173–290, 2007.

[137] H.M. Senn and W. Thiel. Qm/mm methods for biomolecular systems. *Angew. Chem. Int. Ed*, 48:1198–1229, 2009.

[138] A. Warshel and M. Levitt. Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *Journal of molecular biology*, 103(2):227–249, 1976.

[139] Carol V Robinson, Andrej Sali, and Wolfgang Baumeister. The molecular sociology of the cell. *Nature*, 450(7172):973–982, 2007.

[140] Laszlo B Kish. End of moore's law: thermal (noise) death of integration in micro and nano electronics. *Physics Letters A*, 305(3):144–149, 2002.

[141] Johannes Grotendorst, Norbert Attig, Stefan Blügel, and Dominik Marx. Multiscale simulation methods in molecular sciences. *Lecture Notes, NIC Series*, 42, 2009.

[142] Godehard Sutmann. Classical molecular dynamics. *Quantum Simulations of Complex Many-Body Systems: From Theory to Algorithms*, 10:211–254, 2002.

[143] Michele Vendruscolo and Christopher M Dobson. Protein dynamics: Moore's law in molecular biology. *Current Biology*, 21(2):R68–R70, 2011.

[144] KY Sanbonmatsu and C-S Tung. High performance computing in biology: multimillion atom simulations of nanoscale systems. *Journal of structural biology*, 157(3):470–480, 2007.

[145] KY Sanbonmatsu and CS Tung. Large-scale simulations of the ribosome: a new landmark in computational biology. In *Journal of Physics: Conference Series*, volume 46, page 334. IOP Publishing, 2006.

[146] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21:1087, 1953.

[147] BJ Alder and TEf Wainwright. Phase transition for a hard sphere system. *The Journal of Chemical Physics*, 27(5):1208–1209, 1957.

[148] A Rahman. Correlations in the motion of atoms in liquid argon. *phys. Rev*, 136(2A):405–411, 1964.

[149] Aneesur Rahman and Frank H Stillinger. Molecular dynamics study of liquid water. *The Journal of Chemical Physics*, 55:3336, 1971.

[150] J Andrew McCammon. Dynamics of folded proteins. *Nature*, 267:16, 1977.

[151] WF Van Gunsteren and M Karplus. Protein dynamics in solution and in a crystalline environment: a molecular dynamics study. *Biochemistry*, 21(10):2259–2274, 1982.

[152] William E Harte, S Swaminathan, and David L Beveridge. Molecular dynamics of hiv-1 protease. *Proteins: Structure, Function, and Bioinformatics*, 13(3):175–194, 1992.

[153] Yong Duan and Peter A Kollman. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science*, 282(5389):740–744, 1998.

[154] Mu Gao, David Craig, Viola Vogel, and Klaus Schulten. Identifying unfolding intermediates of fn-iii¡ sub¿ 10¡/sub¿ by steered molecular dynamics. *Journal of molecular biology*, 323(5):939–950, 2002.

[155] D Peter Tieleman. The molecular basis of electroporation. *BMC biochemistry*, 5(1):10, 2004.

[156] Peter L Freddolino, Feng Liu, Martin Gruebele, and Klaus Schulten. Ten-microsecond molecular dynamics simulation of a fast-folding ww domain. *Biophys J*, 94(10):L75–L77, May 2008.

[157] Vincent A Voelz, Gregory R Bowman, Kyle Beauchamp, and Vijay S Pande. Molecular simulation of ab initio protein folding for a millisecond folder ntl9 (1- 39). *Journal of the American Chemical Society*, 132(5):1526–1528, 2010.

[158] David E Shaw, Paul Maragakis, Kresten Lindorff-Larsen, Stefano Piana, Ron O Dror, Michael P Eastwood, Joseph A Bank, John M Jumper, John K Salmon, Yibing Shan, et al. Atomic-level characterization of the structural dynamics of proteins. *Science*, 330(6002):341–346, 2010.

[159] Gordon E Moore et al. Cramming more components onto integrated circuits, 1965.

[160] Michael Kanellos. Moores law to roll on for another decade. *CNET News. com*, 2003.

[161] Martin Hilbert and Priscila López. The worlds technological capacity to store, communicate, and compute information. *Science*, 332(6025):60–65, 2011.

[162] David L Donoho et al. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture*, pages 1–32, 2000.

[163] Robert Clarke, Habtom W Ressom, Antai Wang, Jianhua Xuan, Minetta C Liu, Edmund A Gehan, and Yue Wang. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nature Reviews Cancer*, 8(1):37–49, 2008.

[164] P. Bremaud. *Markov chains: Gibbs fields, Monte Carlo simulation and queues.* Springer, 1999.

[165] G.R. Bowman, K.A. Beauchamp, G. Boxer, and V.S. Pande. Progress and challenges in the automated construction of Markov state models for full protein systems. *The Journal of chemical physics*, 131:124101, 2009.

[166] Nina Singhal, Christopher D. Snow, and Vijay S. Pande. Using path sampling to build better markovian state models: Predicting the folding rate and mechanism of a tryptophan zipper beta hairpin. *The Journal of Chemical Physics*, 121(1):415–425, 2004.

[167] Nina Singhal and Vijay S. Pande. Error analysis and efficient sampling in markovian state models for molecular dynamics. *The Journal of Chemical Physics*, 123(20):204909, 2005.

[168] N.S. Hinrichs and V.S. Pande. Calculation of the distribution of eigenvalues and eigenvectors in Markovian state models for molecular dynamics. *The Journal of Chemical Physics*, 126:244101, 2007.

[169] S.P. Elmer, S. Park, and V.S. Pande. Foldamer dynamics expressed via Markov state models. I. Explicit solvent molecular-dynamics simulations in acetonitrile, chloroform, methanol, and water. *The Journal of Chemical Physics*, 123:114902, 2005.

[170] I. Buch, T. Giorgino, and G. De Fabritiis. Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proceedings of the National Academy of Sciences*, 108(25):10184, 2011.

[171] Vijay S Pande, Kyle Beauchamp, and Gregory R Bowman. Everything you wanted to know about markov state models but were afraid to ask. *Methods*, 52(1):99–105, 2010.

[172] Xuhui Huang, Yuan Yao, Gregory R Bowman, Jian Sun, Leonidas J Guibas, Gunnar Carlsson, and Vijay S Pande. Constructing multi-resolution markov state models (msms) to elucidate rna hairpin folding mechanisms. In *Pac. Symp. Biocomput*, volume 15, pages 228–239. World Scientific, 2010.

[173] W.C. Swope, J.W. Pitera, and F. Suits. Describing Protein Folding Kinetics by Molecular Dynamics Simulations. 1. Theory. *J. Phys. Chem. B*, 108(21):6571–6581, 2004.

[174] W.C. Swope, J.W. Pitera, F. Suits, M. Pitman, M. Eleftheriou, B.G. Fitch, R.S. Germain, A. Rayshubskiy, TJC Ward, Y. Zhestkov, et al. Describing Protein Folding Kinetics by Molecular Dynamics Simulations. 2. Example Applications to Alanine Dipeptide and a b-Hairpin Peptide. *J. Phys. Chem. B*, 108(21):6582–6594, 2004.

[175] R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2010. ISBN 3-900051-07-0.

[176] Bernard R Brooks, Charles L Brooks, Alexander D MacKerell, Lennart Nilsson, RJ Petrella, Benoît Roux, Youngdo Won, Georgios Archontis, Christoph Bartels, Stefan Boresch, et al. Charmm: the biomolecular simulation program. *Journal of computational chemistry*, 30(10):1545–1614, 2009.

[177] Berk Hess, Carsten Kutzner, David van der Spoel, and Erik Lindahl. Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of chemical theory and computation*, 4(3):435–447, 2008.

[178] MATLAB. *version 7.10.0 (R2010a).* The MathWorks Inc., Natick, Massachusetts, 2010.

[179] John W Eaton. Octave: Past, present and future. In *Proceedings of the 2nd International Workshop on Distributed Statistical Computing*, 2001.

[180] Kurt Hornik, Friedrich Leisch, and Achim Zeileis. Ten years of octave recent developments and plans for the future. *Proc. DSC 2004*, 2004.

[181] Edward Walker. The real cost of a cpu hour. *Computer*, 42(4):35–41, 2009.

[182] World Health Organization et al. Malaria funding and resource utilization: the first decade of roll back malaria. *Roll Back Malaria Progress and Impact Series*, 1, 2010.

[183] V. Arantes and J.N. Saddler. Cellulose accessibility limits the effectiveness of minimum cellulase loading on the efficient hydrolysis of pretreated lignocellulosic substrates. *Biotechnology for biofuels*, 4(1):1–17, 2011.

[184] R.P. Chandra, A.R. Esteghlalian, and J.N. Saddler. Assessing substrate accessibility to enzymatic hydrolysis by cellulases. *Characterization of Lignocellulosic Materials*, pages 60–80, 2009.

[185] I. Kataeva, G. Guglielmi, and P. Bguin. Interaction between clostridium thermocellum endoglucanase celd and polypeptides derived from the cellulosome-integrating protein cipa: stoichiometry and cellulolytic activity of the complexes. *Biochem J*, 326 (Pt 2):617–624, Sep 1997.

[186] T. Yui, H. Shiiba, Y. Tsutsumi, S. Hayashi, T. Miyata, and F. Hirata. Systematic Docking Study of the Carbohydrate Binding Module Protein of Cel7A with the Cellulose Iα Crystal Model. *J. Phys. Chem. B*, 114(1):49–58, 2010.

[187] G. Carrard, A. Koivula, H. Sderlund, and P. Bguin. Cellulose-binding domains promote hydrolysis of different sites on crystalline cellulose. *Proc Natl Acad Sci U S A*, 97(19):10342–10347, Sep 2000.

[188] T. Reinikainen, O. Teleman, and T.T. Teeri. Effects of pH and high ionic strength on the adsorption and activity of native and mutated cellobiohydrolase I from Trichoderma reesei. *Proteins: Structure, Function, and Bioinformatics*, 22(4):392–403, 1995.

[189] A. L. Creagh, E. Ong, E. Jervis, D. G. Kilburn, and C. A. Haynes. Binding of the cellulose-binding domain of exoglucanase cex from cellulomonas fimi to insoluble microcrystalline cellulose is entropically driven. *Proc Natl Acad Sci U S A*, 93(22):12229–12234, Oct 1996.

[190] A.B. Boraston. The interaction of carbohydrate-binding modules with insoluble non-crystalline cellulose is enthalpically driven. *Biochemical Journal*, 385(Pt 2):479, 2005.

[191] John David Jackson. *Classical Electrodynamics Third Edition*. Wiley, 1998.

[192] K. Kipper, P. Väljamäe, and G. Johansson. Processive action of cellobiohydrolase cel7a from trichoderma reesei is revealed as burstkinetics on fluorescent polymeric model substrates. *Biochemical Journal*, 385(Pt 2):527, 2005.

[193] L.R. Lynd, P.J. Weimer, W.H. Van Zyl, and I.S. Pretorius. Microbial cellulose utilization: fundamentals and biotechnology. *Microbiology and molecular biology reviews*, 66(3):506–577, 2002.

[194] A. Nutt, V. Sild, G. Pettersson, and G. Johansson. Progress curves. *European Journal of Biochemistry*, 258(1):200–206, 1998.

[195] Gregg T Beckham, Yannick J Bomble, Edward A Bayer, Michael E Himmel, and Michael F Crowley. Applications of computational science for understanding enzymatic deconstruction of cellulose. *Curr Opin Biotechnol*, 22(2):231–238, Apr 2011.

[196] G.T. Beckham, Y.J. Bomble, J.F. Matthews, C.B. Taylor, M.G. Resch, J.M. Yarbrough, S.R. Decker, L. Bu, X. Zhao, C. McCabe, et al. The o-glycosylated linker from the trichoderma reesei family 7 cellulase is a flexible, disordered protein. *Biophysical journal*, 99(11):3773–3781, 2010.

[197] Lintao Bu, Gregg T Beckham, Michael F Crowley, Christopher H Chang, James F Matthews, Yannick J Bomble, William S Adney, Michael E Himmel, and Mark R Nimlos. The energy landscape for the interaction of the family 1 carbohydrate-binding module and the cellulose surface is altered by hydrolyzed glycosidic bonds. *J Phys Chem B*, 113(31):10994–11002, Aug 2009.

[198] Chandrika Mulakala and Peter J Reilly. Hypocrea jecorina (trichoderma reesei) cel7a as a molecular machine: A docking study. *Proteins*, 60(4):598–605, Sep 2005.

[199] B.J. Berne, J.D. Weeks, and R. Zhou. Dewetting and hydrophobic interaction in physical and biological systems. *Physical Chemistry*, 60(1):85, 2009.

[200] Ken A Dill, Thomas M Truskett, Vojko Vlachy, and Barbara Hribar-Lee. Modeling water, the hydrophobic effect, and ion solvation. *Annu Rev Biophys Biomol Struct*, 34:173–199, 2005.

[201] Lawrence R Pratt. Molecular theory of hydrophobic effects: "she is too mean to have her name repeated.". *Annu Rev Phys Chem*, 53:409–436, 2002.

[202] Ruhong Zhou, Xuhui Huang, Claudio J Margulis, and Bruce J Berne. Hydrophobic collapse in multidomain protein folding. *Science*, 305(5690):1605–1609, 2004.

[203] I. Daidone, M.B. Ulmschneider, A. Di Nola, A. Amadei, and J.C. Smith. Dehydration-driven solvent exposure of hydrophobic surfaces as a driving force in peptide folding. *Proceedings of the National Academy of Sciences*, 104(39):15230, 2007.

[204] G..R. Pettersson, M. Linder, T. Reinikainen, T. Drakenberg, M.L. Mattinen, A. Annila, M. Kontteli, G. Lindeberg, and J. Ståhlberg. Identification of functionally important amino acids in the cellulose-binding domain of Trichoderma reesei cellobiohydrolase I. *Protein Science*, 4(6):1056–1064, 1995.

[205] G.T. Beckham, J.F. Matthews, Y.J. Bomble, L. Bu, W.S. Adney, M.E. Himmel, M.R. Nimlos, and M.F. Crowley. Identification of amino acids responsible for processivity in a family 1 carbohydrate-binding module from a fungal cellulase. *The Journal of Physical Chemistry B*, 114(3):1447–1453, 2010.

[206] Thomas Neusius, Isabella Daidone, Igor M Sokolov, and Jeremy C Smith. Subdiffusion in peptides originates from the fractal-like structure of configuration space. *Physical review letters*, 100(18):188103, 2008.

[207] Thomas Neusius, Igor M Sokolov, and Jeremy C Smith. Subdiffusion in time-averaged, confined random walks. *Physical Review E*, 80(1):011109, 2009.

[208] Thomas Neusius, Isabella Daidone, Igor M Sokolov, and Jeremy C Smith. Configurational subdiffusion of peptides: A network study. *Physical Review E*, 83(2):021902, 2011.

[209] J. Wohlert and L.A. Berglund. A coarse-grained model for molecular dynamics simulations of native cellulose. *Journal of Chemical Theory and Computation*, 2011.

[210] E.J. Jervis, C.A. Haynes, and D.G. Kilburn. Surface diffusion of cellulases and their isolated binding domains on cellulose. *Journal of Biological Chemistry*, 272(38):24016, 1997.

[211] Mark R Nimlos, Gregg T Beckham, James F Matthews, Lintao Bu, Michael E Himmel, and Michael F Crowley. Binding preferences, surface attachment, diffusivity, and orientation of a family 1 carbohydrate-binding module on cellulose. *Journal of Biological Chemistry*, 287(24):20603–20612, 2012.

[212] K. Igarashi, A. Koivula, M. Wada, S. Kimura, M. Penttil, and M. Samejima. High speed atomic force microscopy visualizes processive movement of Trichoderma reesei cellobiohydrolase I on crystalline cellulose. *Journal of Biological Chemistry*, 284(52):36186, 2009.

[213] F. Eisenhaber, P. Lijnzaad, P. Argos, C. Sander, and M. Scharf. The double cubic lattice method: efficient approaches to numerical integration of surface area and volume and to dot surface contouring of molecular assemblies. *Journal of Computational Chemistry*, 16(3):273–284, 1995.

[214] H. Shen, M. Schmuck, I. Pilz, NR Gilkes, DG Kilburn, RC Miller, and RA Warren. Deletion of the linker connecting the catalytic and cellulose-binding domains of endoglucanase a (cena) of cellulomonas fimi alters its conformation and catalytic activity. *Journal of Biological Chemistry*, 266(17):11335, 1991.

[215] LM Ferreira, A.J. Durrant, J. Hall, G.P. Hazlewood, and HJ Gilbert. Spatial separation of protein domains is not necessary for catalytic activity or substrate binding in a xylanase. *Biochemical Journal*, 269(1):261, 1990.

[216] Hans Christian Öttinger. *Beyond equilibrium thermodynamics.* Wiley-Interscience, 2005.

[217] S. Cheng, M. Cetinkaya, and F. Gräter. How sequence determines elasticity of disordered proteins. *Biophysical journal*, 99(12):3863–3869, 2010.

[218] C. Dicko, D. Porter, J. Bond, J.M. Kenney, and F. Vollrath. Structural disorder in silk proteins reveals the emergence of elastomericity. *Biomacromolecules*, 9(1):216–221, 2007.

[219] S. Rauscher, S. Baud, M. Miao, F.W. Keeley, and R. Pomès. Proline and glycine control protein self-organization into elastomeric or amyloid fibrils. *Structure*, 14(11):1667–1676, 2006.

[220] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983.

[221] R.P. Joosten, T.A.H. te Beek, E. Krieger, M.L. Hekkelman, R.W.W. Hooft, R. Schneider, C. Sander, and G. Vriend. A series of pdb related databases for everyday needs. *Nucleic acids research*, 39(suppl 1):D411–D419, 2011.

[222] A. Guinier, G. Fournet, C.B. Walker, and K.L. Yudowitch. *Small-angle scattering of X-rays*, volume 14. Wiley New York, 1955.

[223] PM Abuja, M. Schmuck, I. Pilz, P. Tomme, M. Claeyssens, and H. Esterbauer. Structural and functional domains of cellobiohydrolase i from trichoderma reesei. *European Biophysics Journal*, 15(6):339–342, 1988.

[224] P.M. Abuja, I. Pilz, M. Claeyssens, and P. Tomme. Domain structure of cellobiohydrolase ii as studied by small angle x-ray scattering: close resemblance to cellobiohydrolase i. *Biochemical and biophysical research communications*, 156(1):180–185, 1988.

[225] M. Schmuck, I. Pilz, M. Hayn, and H. Esterbauer. Investigation of cellobiohydrolase from trichoderma reesei by small angle x-ray scattering. *Biotechnology letters*, 8(6):397–402, 1986.

[226] S.V. Pingali, H.M. O'Neill, J. McGaughey, V.S. Urban, C.S. Rempe, L. Petridis, J.C. Smith, B.R. Evans, and W.T. Heller. Small angle neutron scattering reveals ph-dependent conformational changes in trichoderma reesei cellobiohydrolase i implications for enzymatic activity. *Journal of Biological Chemistry*, 286(37):32801–32809, 2011.

[227] Ng, Randles, and Clarke. *Protein Folding Protocols*, volume 350. Humana Press, 2006.

[228] W. Min, G. Luo, B.J. Cherayil, SC Kou, and X.S. Xie. Observation of a power-law memory kernel for fluctuations within a single protein molecule. *Physical review letters*, 94(19):198302, 2005.

[229] W. Min, P. Brian, G. Luo, B.J. Cherayil, SC Kou, and X.S. Xie. Fluctuating enzymes: lessons from single-molecule studies. *Accounts of chemical research*, 38(12):923–931, 2005.

[230] GR Kneller and K. Hinsen. Fractional brownian dynamics in proteins. *The Journal of chemical physics*, 121:10278, 2004.

[231] Deanne W. Sammond, Christina M. Payne, Roman Brunecky, Michael E. Himmel, Michael F. Crowley, and Gregg T. Beckham. Cellulase linkers are optimized based on domain type and function: Insights from sequence analysis, biophysical measurements, and molecular simulation. *PLoS ONE*, 7(11):e48615, 11 2012.

[232] Hiroshi Akima. A method of bivariate interpolation and smooth surface fitting for irregularly distributed data points. *ACM Trans. Math. Softw.*, 4(2):148–159, June 1978.

[233] Hiroshi Akima. Algorithm 761: Scattered-data surface fitting that has the accuracy of a cubic polynomial. *ACM Trans. Math. Softw.*, 22(3):362–371, September 1996.

[234] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics).* Springer, 2007.

[235] Yaser S. Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin. *Learning From Data.* AMLBook, 2012.

[236] Jorge J Moré. The levenberg-marquardt algorithm: implementation and theory. In *Numerical analysis*, pages 105–116. Springer, 1978.

[237] Gerhard Hummer, Shekhar Garde, Angel E García, Andrew Pohorille, and Lawrence R Pratt. An information theory model of hydrophobic interactions. *Proceedings of the National Academy of Sciences*, 93(17):8951–8955, 1996.

[238] Shekhar Garde, Gerhard Hummer, Angel E García, Michael E Paulaitis, and Lawrence R Pratt. Origin of entropy convergence in hydrophobic hydration and protein folding. *Physical review letters*, 77(24):4966–4968, 1996.

[239] F. H. Stillinger. *J. Solution Chem.*, 2:141–158, 1973.

[240] Ingrid Wagner and Hans Musso. New naturally occurring amino acids. *Angewandte Chemie International Edition in English*, 22(11):816–828, 1983.

[241] Jack Kyte and Russell F Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology*, 157(1):105–132, 1982.

[242] Wikipedia. Amino acid — wikipedia, the free encyclopedia, 2013. [Online; accessed 16-May-2013].