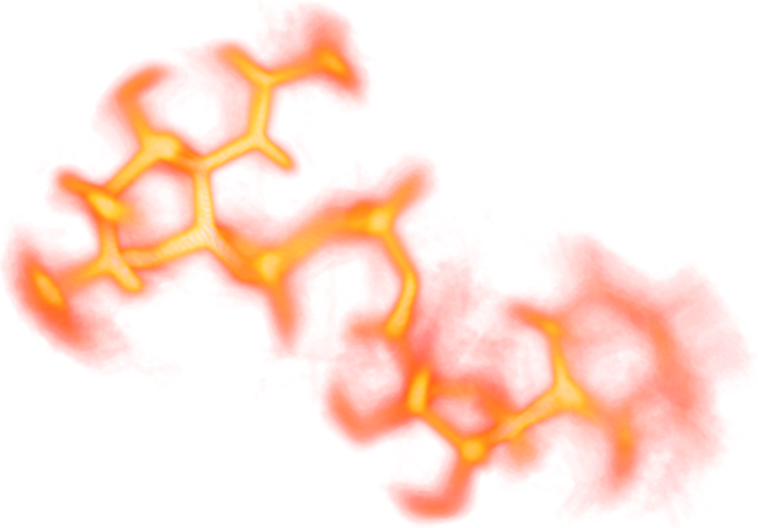
Master's Thesis
in the Field of Bioinformatics

Conformational Studies of UDP-N-Acetyl-Glucosamine in Environments of Increasing Complexity

submitted by
Martin Held, June 7, 2007



Freie Universität Berlin

CHARITÉ
UNIVERSITÄTSMEDIZIN BERLIN

Betreuer: Prof. Dr. Christof Schütte
Prof. Dr. Werner Reutter

I Martin Held declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institute of tertiary education. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given in the bibliography.

Cover image produced using Amira [**?**]

Contents

List of Figures

List of Tables

1 Introduction

The role of computational biology has shown a steady increase over the past decades. With the advancement in computational power and analysis methodology it became possible to evaluate, compare and characterize large experiments of biological systems. This in turn made it possible to build models of biological processes that are accessible for analysis and simulations. Possible simulation scenarios in that context include, e.g., cellular signaling pharmacokinetics or molecular dynamics (MD). In particular, molecular dynamics simulations have proven eminently successful in the area of analyzing protein folding processes, conformational studies of bioactive compounds and virtual drug design.

In this thesis we perform, to our knowledge, the first MD based conformational study of the key substrate of the sialic acid pathway. Sialic acids play an important role in various biological processes, e.g., immune response, tumor metastasis or inflammatory reactions. Hence, an understanding of the key substrate that accounts for the limiting step in the pathway is worthwhile to pave the way for designing potential inhibitors making it possible to interfere the sialic acid synthesis and thus develop treatments for corresponding diseases. The key reaction is in human conducted by the UDP-GlcNAc-2-Epimerase/ManNAc-Kinase (GNE) [Hinderlich et al., 1997, Stäsche et al., 1997, Effertz et al., 1999]. It carries out the, by feedback inhibition, regulated step of epimerizing UDP-GlcNAc to ManNAc. For a better understanding of the ligand behavior we study the general conformational dynamics of UDP-N-acetyl-glucosamine (UDP-GlcNAc). We compare its characteristics in vacuum and in water environment by performing and analyzing molecular dynamics simulations of it.

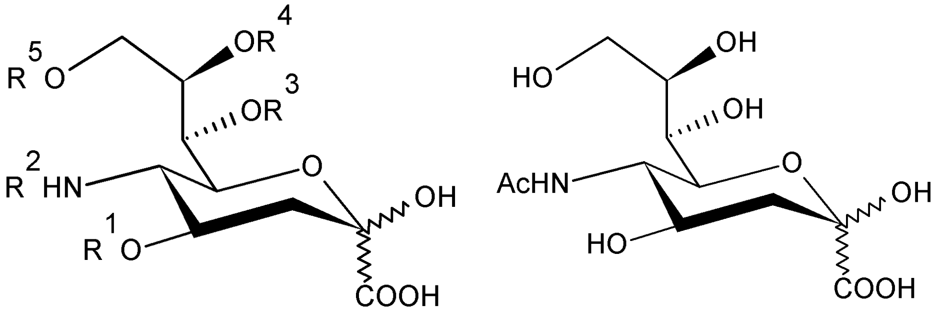
As molecular dynamics simulations are indeed widely-used for the study of peptides or nucleic acids, it can be rather elaborate to use them with unusual molecules due to missing parameters needed by MD programs to describe a molecule. UDP-GlcNAc is a representative of such an uncommon molecule as its atom composition is quite unusual concerning classical MD targets. Therefore it was necessary to determine the missing parameters before MD simulations could be carried out. Fortunately, the missing parameters after could be obtained after surveying the literature and combining two located parameter sets, enabling the simulation of UDP-GlcNAc in different environments. The computational costs of simulations depend on system size and simulated time range. In order to dilute this effect, a distributed computing approach that allows to simulate different conformers of UDP-GlcNAc in parallel was applied. The simulations resulted in trajectories of UDP-GlcNAc in vacuum and water environment which where analyzed in respect of metastable conformations, i.e., geometrical large scale structures that persit for long periods of time but rarely switch to other conformations. Within this interpretation of metastability it has been a promising idea to describe the metastable dynamics of the system by means of a Markov chain which state space represents the possible conformations and the transition matrix models the "flipping-dynamics" between conformations. Furthermore, this understanding of metastability allows for the application of a hidden Markov model (HMM) based approach to discretize the state space and the application of Perron Cluster Cluster Analysis (PCCA), a method exploiting the spectral properties of the transition matrix, to identify metastable conformations of UDP-GlcNAc.

Currently, the correct binding orientation of UDP-GlcNAc to the UDP-GlcNAc-2-Epimerase is unknown, which renders it difficult to infer knowledge about the reaction mechanism and crucial amino acids involved in the reaction, hence prohibiting a rational inhibitor design. To remedy this situation we modelled several potential binding conformations of UDP-GlcNAc into the binding pocket of UDP-GlcNAc-2-Epimerase - *E.Coli* (the E.Coli homologue of the enzyme was chosen, because for the human homologue not crystallographic 3D data is currently available). The modelled binders were evaluated and tested by means of MD simulations to check if they are stable. To make a conjecture about which binder orientation is the correct one, the generated water trajectory data were utilized by screening them for the structurally known UDP orientation.

This thesis is structured as follows, the first part is concerned with the biological relevance of sialic acids, their synthesis pathway and the resulting importance of UDP-GlcNAc. This section is followed by a synopsis about molecular dynamics simulations in general, including a description of the parameterization of UDP-GlcNAc. The third section explains in detail the performed simulations and their parameters as well as the procedure used to distribute the computing jobs.

The fourth section elaborates on the hidden Markov model (HMM) approach used in conjunction with Perron Cluster Cluster Analysis (PCCA) to identify metastable sets from the simulation trajectories. It also outlines the procedures used to examine potential binding structures of UDP-GlcNAc.

The simulation and conformational analysis results are presented in section 5. Finally, section 6 discusses the findings made and in the last section conclusions are drawn and an outlook for further work is given.



(a) General Sialic Acid Structure – R^2 can be replaced by acetyl or glycolyl groups; R^1, R^3,R^4 and R^5 by acetyl-, lactoyl-, methyl-, sulfonyl- or phosphonyl groups

(b) Neu5Ac – most frequent sialic acid, all hydroxyl groups are unmodified

Figure 1

2 Background

2.1 Sialic Acids

Sialic acids have first been identified by E. Klenk [1935] in brain tissue and by G. Blix [1936] in submaxillary mucin. They form an integral part of glycoconjugates in *deuterostome*[1] organisms. Chemically sialic acids belong to the family of acidic amino sugars. The C-2 position always carries a carboxyl group and the C-5 position an amino group. The most frequent sialic acid, and also precursor of most of the other, is N-Acetylneuraminic Acid (Neu5Ac), it contains an acetylated amino group at C-5 (Figure 1).

Due to the high number of possible substitutes 1 a)), sialic acids exhibit a great diversity in structure and function. Their appearance and abundance varies greatly in different species and tissues [Varki, 1993]. The specific distribution is thereby controlled by the activity of a wide range of sialyltransferases [Paulson et al., 1989]. In vertebrates sialic acids occur on glycoproteins (N- or O-linked) as well as in gangliosides, where they are usually found in terminal position. This is in contrast to bacteria, where they are normally located in the inner part of polysaccharide chains in bacterial walls.

[1] Greek: "second mouth" – superphylum of animals, refers to an important developmental feature unique to this group

Sialic acids are important for various biological functions. Due to their negative charge and terminal position in glycoconjugates they, one the one hand, contribute to repulsion effects between cells or cells and extracellular matrix. On the other hand, they are essential for adhesion processes mediated by sialic acid binding lectins. Biological processes involving such forces are, e.g., cell migration, tissue development, inflammatory reactions, tumor growth or the development of metastasis. A well-studied example for cell migration is the "rolling" and migration of leukocytes into inflamed tissue. It is induced by selectins on the leukocyte that recognize specific sialic acids [Lasky, 1995] at the center of inflammation. An additional characteristic of sialic acids is the ability to serve as recognition domain for viruses, parasites and toxins [Schauer, 1985, Varki, 1992, Karlsson, 1995]. The specificity of the influence A virus family , for example, is determined by the sialisation pattern of the host cell. Depending on the virus type it recognizes and infects cells from specific organsims (e.g. human, chicken, pig) [Ito et al., 1998]. Only if there is a change (mutation or horizontal gene transfer) in the sialic acid binding domain of the virus, it is capable of infection "foreign" hosts. Considering the bird flu virus, it is exactly such an event everybody is currently concerned about.

Furthermore it has been observed that sialic acids protect parasite micro organisms and certain cells from recognition by the immune system. For the parasite *trypanosoma cruzi* it has been found that it copies the sialisation pattern of his host and is thus protected from immune system response [Colli, 1993, Tomlinson et al., 1994]. During embryogenesis sialic acids protect embryonic cells from degradation by the mother immune system [Schauer, 1985]. Apart from protection, sialic acids can also trigger a specific immune response, e.g., as blood group antigenes [Pilatte et al., 1993].

Sialic acids also play an important role in cancer biology. An increased sialic acid concentration on tumor cell surfaces often indicates an increased tumor malignity [Fogel et al., 1983, Hakomori, 1989, Bresalier et al., 1990, Bhavanandan, 1991, Sawada et al., 1994]. Similar to leukocyte migration, sialic acid mediated adhesion can increase the invasiveness of tumor cells [Kageshita et al., 1995]. The above mentioned immune system protection effect sialic acids can have is also be found in various cancers [Dennis and Lafert, 1985]. For tumor diagnostics sialic acids can serve as well, e.g., for melanoma detection $\mathrm{Neu9Ac_2}$ can be utilized [Fahr and Schauer, 2001]

2.1.1 Sialic Acid Synthesis

Sialic acids are produced via the sialic acid synthesis pathway. It consists of several steps leading to CMP-Neu5Ac, the common sialic acid precursor. The initial substrate of the pathway is the nucleotide sugar UDP-GlcNAc. UDP-GlcNAc has a high abundance in cells, ranging from 100-200 μM [?]. It is synthesized in a part of the general amino sugar metabolism. In its role in the sialic acid pathway it marks the beginning of the metabolism. In this initial step UDP-GlcNAc is processed by the bifunctional enzyme UDP-GlcNAc-2-Epimerase/ManNAc-Kinase [Hinderlich et al., 1997, Stäsche et al., 1997, Effertz et al., 1999]. It is the feedback inhibited (by CMP-Neu5Ac) limiting step of the pathway. The UDP-GlcNAc-Epimerase domain eliminates UDP and epimerises the N-Acetyl group at the C-2 position to ManNAc in the first reaction step. In the second step, the ManNAc-Kinase domain phosphorylates ManNAc to ManNAc-6-P.

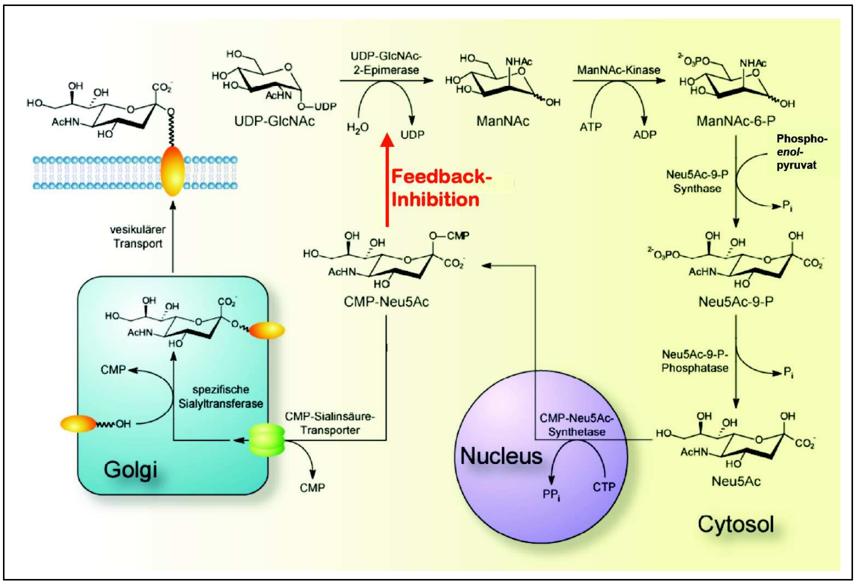


Figure 2: Sialic Acid Synthesis Pathway

The next steps comprise condensation, phosphate elimination and activation in the nucleus (as depicted in Figure 2), such that the final product is CMP-Neu5Ac.

Being able to regulate the sialic acid synthesis, might permit the development of treatments for associated diseases. For this reason it is necessary to control the limiting step of the reaction, the conversion of UDP-GlcNAc to ManNAc. To do so, the design of an UDP-GlcNAc analogon which can influence the efficiency of the limiting reaction is desireable.

In this work, we undertake a first step into the direction of finding such an inhibitor by characterizing the conformational dynamics and potential binding conformations of UDP-GlcNAc by means of molecular dynamics simulations.

2.2 Molecular Dynamic Simulations

Molecular dynamic (MD) simulations have become a principal tool in the theoretical study of biomolecular systems. They are used to calculate the time dependent behavior of molecules, based on known laws of physics. MD simulations provides insights on fluctuations and conformational changes of proteins and nucleic acids. They are used to investgate structure, dynamics and thermodynamics of biological molecules and their complexes. MD can be understood as an interface acting as a bridge between laboratory experiments and theory.

Molecular dynamics simulations have first been performed by Alder and Wainwright [1957], they studied the interactions of hard spheres and could contribute important insights about the behavior of simple liquids. Two decades later, Stillinger and Rahman [1974] performed the first simulation of a real system by simulating liquid water and in 1977 the first protein simulation was carried out, when the pancreatic trypsin inhibitor (BPTI) was simulated by [McCammon et al., 1977]. Today, MD simulations of solvated proteins, Protein-DNA complexes or lipid systems routinely appear in the literature.

By running MD simulations, information about a system at the microscopic level is generated, i.e., information about atom positions and atom velocities over time. However, often one is rather interested in the macroscopic properties (radial distributions, molecule conformations, thermodynamic properties) of a system which can be observed by real experiments. To relate both scales statistical mechanics is employed, it provides the mathematical formalism that relates macroscopic properties, which can be observed in experiments, to microscopic properties of the system, e.g., the distribution and motion single atoms. Statistical mechanics achieves this by introducing the fundermental concept of an statistical ensembles. An ensemble can be understood as a collection of copies of the same molecular system, in which all copies can be in a different microscopic state but, on average, have all the same macroscopic property (temperature, pressure). For this relationship to be valid the fundermental axiom of ”ergodicity” must hold for the system under consideration. The ”ergodic hypothesis” states: If a system is allowed to evolve in time indefinitely, it will definitely pass through all its possible states. The time average of the system will therefore in consequence correspond to its ensemble average. For MD simulations this means that the total simulation time has to be long enough such that the system visits all its possible states. Only the averages of such simulations can be reasonably interpreted as ensemble averages.

Classical molecular dynamic simulations are based on Newton's second law of motion.

$$F_i = m_i a_i$$

Given the force acting on a particle i it is possible to determine it acceleration and by integrating the corresponding equations of motion it yields a trajectory that describes positions, velocities and accelerations of the particles as they vary in time. In the molecular setting forces can also be expressed as

$$F_i = -grad(E(i)),$$

where $grad(E(i))$ denotes the gradient of the potential energy of particle i. If the potential energy of an atom would be known, the force acting on it would be simply given by the derivative of the energy. Unfortunately, this potential energy is in general not known exactly. To be nevertheless able to calculate forces and consequently to perform molecular dynamic simulations, this energy can approximated by using empirical force fields. Such force fields approximate the potential energy by using empirically determined functions for different parts of the total potential. The total potential is usually split up into potentials of bonded interactions and non-bonded interactions.

$$\begin{aligned} E_{total} &= E_{bonded} + E_{non-bonded} \\ E_{bonded} &= E_{bonds} + E_{angles} + E_{dihedrals} \\ E_{non-bonded} &= E_{van\ der\ Waals} + E_{electrostatic} \end{aligned}$$

Bonded interactions comprise bond, bond-angle and dihedral interactions, whereas non-bonded comprise *van der Waals* and electrostatic interactions. Figure 1 depicts the different potential types and corresponding functions used to model them.

Considering the depicted functions for each potential type, an inherent limitation of empirical force fields becomes apparent. Each function needs several parameters to correctly reproduce the position dependend potential. Due to different physical properties, these parameters vary for different atom compositions, e.g., the stretching two bound carbon atoms needs a different force than stretching two hydrogen atoms apart. The determination of such parameters is in general a non trivial task. In general, they are assigned by utilizing *ab initio* quantum calculations or tuned by comparing experimental observations to force field calculations. This labour intensive task has been performed for many common systems such as amino acids or nucleic acids and has been stored in

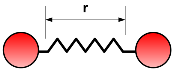
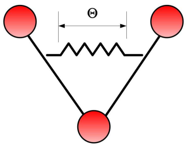
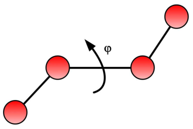
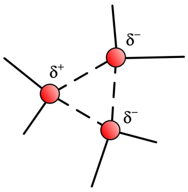
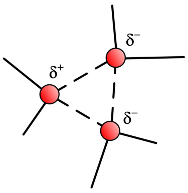
$E_{bonds} = \sum_{bonds} K_r(r - r_{eq})^2$	
$E_{angles} = \sum_{angles} K_\Theta(\Theta - \Theta_{eq})^2$	
$E_{dihedrals} = \sum_{dihedrals} \frac{V_n}{2}(1+cos(n\phi-\gamma))$	
$E_{vdW} = \sum_{i<j}^{atoms} \frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}}$	
$E_{electrostatic} = \sum_{i<j}^{atoms} \frac{q_i q_j}{\epsilon R_{ij}}$	

Table 1: Force Field Terms as Used by the Simulation Package Amber 8

parameters sets available in various molecular simulation software packages. However, for systems with an atom composition not contained in these data sets, it can by rather difficult to obtain suitable parameters, e.g., in the present case of the UDP-GlcNAc simulation. See section 2.3 for more details.

2.2.1 Periodic Boundary Conditions

The overall system size that can be simulated by computational methods is limited. Nowadays, it is in the range of a million atoms, a fairly small number when compared to realistic system sizes in the molar range ($N_A = 6,02 * 10^{23}$). With this big discrepancy a problem comes along concerning the size of the systems surface. For small systems the surface to size ratio is much bigger than for larger systems. This means that in molecular simulations surface effects become more important than they actually should. To circumvent this, periodic boundary conditions (PBC) have been introduced. When PBC are used the particles of a system are enclosed in a box and this box is replicated to infinity in all three carthesian directions, completely filling the space. The particle coordinates are then defined as

$$\mathbf{r} = (ix + jy + kz) \qquad i, j, k \in [-\infty, +\infty],$$

where x,y,z correspond to vectors determining the box edges. If a particle moves, all its infinite copies move as well, when it lefts the box it is copied back into it at the corresponding position (see Figure 3).

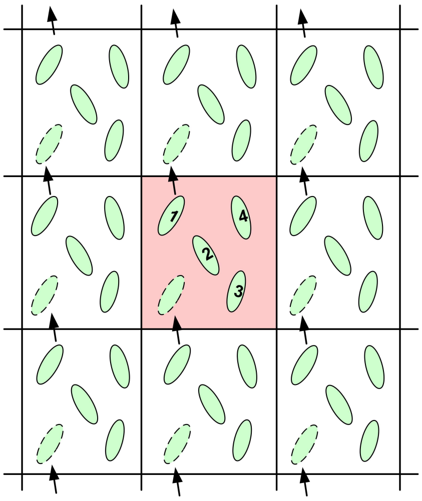


Figure 3: Periodic Boundary Conditions

Particles can theoretically interact with all other particles, but this would result in an infinite amount of necessary computation. Hence, in practice, only interactions with particles within a certain cutoff are computed. This introduces a slight error, but is permissible since non-bonded interactions are usually only short ranged.

2.2.2 Experimental Conditions

Molecular Dynamic simulations as described above yield trajectories describing an NVE ensemble (constant number of particles, constand volume and constand energy), because Newton's second law of motion ($F = ma$) obeys energy conservation. Unfortunately, the NVE setting is difficult to accomplish in experimental practice. It is often impossible to completely isolate the system such as the total energy is conserved, it is rather common that experiments are carried out under constant temperature and constant pressure conditions (Isothermal-Isobaric Ensemble NPT).

To enable MD simulations to produce trajectories of this ensemble type various extentions of the basic Newton dynamics have proposed developed. They can be grouped into simple constrained formulations, including stochastically motivated models and more sophisticated techniques involving additional degrees of freedom. The MD program used in the present study uses stochastic methods. It uses Langevin Dynamics [Schlick, 2002] to model the temperature constraint and a Berensen pressure coupling scheme [Berendsen et al., 1984] to account for the pressure condition. Langevin dynamics represents a simple heat bath by accounting for molecular collisions. It achieves this by adding a collision term and a random force vector to the standard potential, resulting in a stochastic differential equation. It can be shown that the resulting kinetic energy of the simulated system then corresponds on average to the target temperature. The Bersensen scheme preserves the pressure by coupling the system to an external "pressure bath" that rescales the particle positions such that the pressure remains constant.

2.3 Modeling of UDP-GlcNAc

UDP-GlcNAc does not belong to a class of molecules classical force fields have been particularly designed and parameterized for. This poses several hurdles on a simulation study of that molecule. First of all, a three dimensional representation of the molecule had to be created. There are several modeling tools available for that purpose, which allow to draw a 2D structural formula and infer a 3D structure from it. However, the resulting 3D structures are often flat and appear not to resemble the native structure. We therefore screened several structural databases for UDPGlcNAc and finally obtained the structure from the Biological Magnetic Resonance Data Bank [Seavey et al., 1991]. It was provides in PDB format and was converted by using *antechamber* to be readable by the used simulation program.

The second hurdle concerned the missing partial charges of the molecule needed to correctly model its electrostatic interactions. The standard way to obtain these charges is to perform several exact quantum chemical calculations to obtain an electrostatic potential field (ESP) for different conformations of the molecule. These fields are then used to fit a partial charge to each atom such the partial charges would optimally resemble the ESP field. This charge fitting procedure is in general a rather elaborate task involving a lot of complex computations. Hence, we decided to rely on already determined charges for UDP-Glc [Petrová et al., 1999] and GlcNAc [Woods Group, 2007] published in the literature. From the UDP-Glc publication we transferred the charges only for the UDP part of the molecule, the charges for GlcNAc were additionally obtained from the Glycam parameter set. A list of finally used partial charges is given in Table 2.

The combination of charges from both sources resulted in a net charge of -1.998 for UDP-GlcNAc, instead of -2 . After a careful investigation, it became apparent that the charge of the O-Atom, linking UDP and GlcNAc, caused this difference. Based on former experiences we decided to scale the charge of this atom such that the netto charge of entire molecule was -2.

Finally, the problem of missing force field parameters had to be solved. Above we outlined the need for specific force field parameters depending on the considered atom composition. Unfortunately, the atom composition of UDP-GlcNAc is rather unusual in respect to classical force field application fields resulting in missing angle and dihedral parameters for the phosphate backbone of the molecule. To resolve this problem we again fall back on published force field parameters determined by Petrová et al. [1999]. After having determined all missing parameters we were able to create an input file for the used Amber simulation package containing all necessary information to perform simulations.

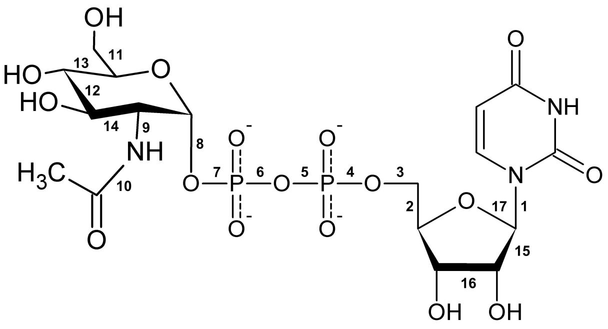


Figure 4: 2D Structural Formula of UDP-GlcNAc with Dihedral Numbering

3 Simulations

3.1 Sampling Problem

Biological processes take place on timescales ranging from a few nanoseconds to several milliseconds or even seconds and they are often determined by corresponding biomolecular conformation changes. However, the time scales accessible by current MD simulations reach barely beyond the microsecond scale. The limiting factor in that respect is determined by the simulated time step size, which must be chosen in a way in a way such that a consistent integration of the equations of motion is possible. When molec-

	Atom	Atom Type	Charge		Atom	Atom Type	Charge
GlcNAc	C	CG	0.06300	Di-Phosphat	O15	OS	-0.42330
	H	HC	0.00000		P	P	1.11280
	H1	HC	0.00000		O5	O2	-0.79670
	H2	HC	0.00000		O10	O2	-0.79670
	C15	C	0.58800		O14	OS	-0.49740
	O9	O	-0.57900	Uracil	C4	CT	0.05580
	N1	N	-0.55200		H7	H1	0.06790
	H18	H	0.23600		H8	H1	0.06790
	C9	CG	0.24500		C8	CT	0.10650
	H11	H1	0.00000		O13	OS	-0.35480
	C12	CG	0.16500		H10	H1	0.11740
	O3	OH	-0.63000		C11	CT	0.20220
	H22	HO	0.41300		O2	OH	-0.65410
	H14	H1	0.00000		H21	HO	0.43760
	C10	CG	0.32200		H13	H1	0.06150
	O1	OH	-0.73300		C13	CT	0.06700
	H20	HO	0.44900		O4	OH	-0.61390
	H12	H1	0.00000		H23	HO	0.41860
	C7	CG	0.24600		H15	H1	0.09720
	C3	CG	0.32800		C14	CT	0.06740
	O	OH	-0.68800		H16	H2	0.18240
	H19	HO	0.42100		N2	N*	-0.04180
	H5	H1	0.00000		C2	CM	-0.11260
	H6	H1	0.00000		H4	H4	0.21880
	H9	H1	0.00000		C1	CM	-0.36350
	O12	OS	-0.56800		H3	HA	0.18110
	C16	CG	0.53770		C5	C	0.59520
	H24	H2	0.00000		O7	O	-0.57610
Di-Phosphat	O16	OS	-0.47630		N	NA	-0.35490
	P1	P	1.09030		H17	H	0.31540
	O6	O2	-0.79280		C6	C	0.46870
	O11	O2	-0.79280		O8	O	-0.54770

Table 2: Atom Types and Partial Charges for UDP-GlcNAc

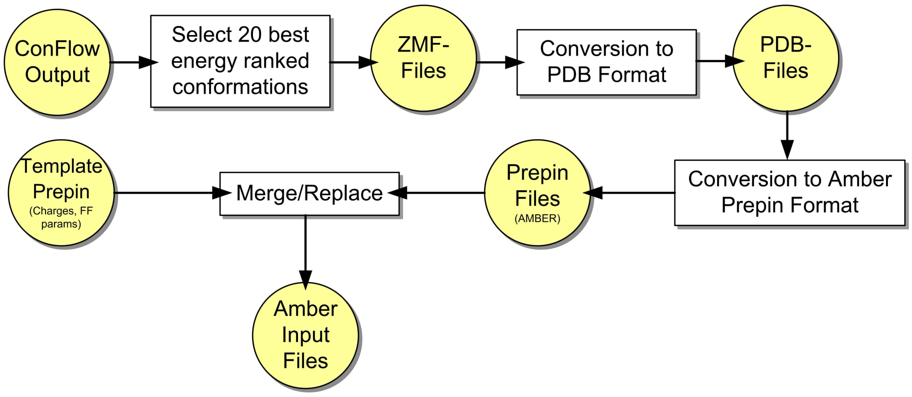

ConFlow Output
Select 20 best energy ranked conformations
ZMF-Files
Conversion to PDB Format
PDB-Files
Conversion to Amber Prepin Format
Prepin Files (AMBER)
Merge/Replace
Template Prepin (Charges, FF params)
Amber Input Files

Figure 5: Generation of Amber Input Files from ConFlow Output

ular systems are concerned this time step is usually in the range of femto second, the timescale H-bond oscillations take place. Given this limitation the obtuseness of MD simulations becomes clear, e.g., generating a 1ns MD simulation involves 1.000.000 calculation steps. This possibly occurring timescale discrepancies can be problematic. It might for example happen that either too few transitions between conformations can be observed to make a valid statistical statement or, even worse, no transitions at all can be observed in the simulation data, which would mean to overlook certain conformations. To counteract this problem several parallel simulations of UDP-GlcNAc were started, each with a different start conformation, assuming that this leads a better sampling of the conformational space and allows for a more reliable description of transitions. To generate the different conformations of UDP-GlcNAc the ConFlow [Meyer et al., 2006] program was used. ConFlow generates conformations of a molecule by a systematic conformational search. Given the ConFlow output we chose to selected the 20 best energy ranking conformations to get a set of diverse start positions. Apart from the improved sampling properties, the 20 conformations enabled us to run simulations of them in parallel, using a distributed computing approach. This approach was based on an agent based scheduling system capable of distributing tasks to various machines. The agents are started independently on different machines polling a database for new simulation tasks. Each time a new task is available it is processed by an agent and the resulting output is written to a central directory. By utilizing this distribution system it was possible to achieve a fairly long total simulation time, generating enough data for a profound analysis.

3.2 Simulation Procedure

To carry out molecular dynamic simulations the Amber 8 [Case et al., 2004] software package was used. The software requires input files defining the structure to simulate and a file containing the simulation parameters. To be able to simulate the generated conformations of UDP-GlcNAc, the output of ConFlow first had to be converted into a format readable by Amber. This was done by manually generating a template file containing all residues, the charges and corresponding force field terms as obtained from the literature. By applying several amber conversion tool the conformers generated by ConFlow could be converted to an Amber readable format and the template file was used to generate the appropriate input files containing the right charges. With the principal input files for each of the 20 conformers the simulations were set up. We run 3 different types of simulations in different environmental conditions (vacuum, water and peptide). The next sections describe the corresponding simulation protocols and parameters.

3.2.1 Vacuo

The first simulation of UDP-GlcNAc were performed *in vacuo* without any solvent present. This setting does not reproduce realistic conditions, however the systems complexity is very low in this case and allows very long simulations. Hence, it provides first valuable insights into the dynamics and flexibility of the ligand. In case of vacuum simulations, the simulation protocol is fairly simple. It starts with minimizing each of the 20 conformers in the given force field to resolve bad particle contacts and to get out of unfavored configurations. As minimization algorithms steepest decent and conjugate gradient are provided by Amber. In the present case we run 1000 steps of steepest decent minimization to approach the minimum quickly, to then obtain a better convergence rate the last 1000 steps are carried out using conjugate gradient. After the system was minimized, it was heated from 0K to 300K within 20ps using Langevin temperature control. This slow heating process is necessary to obtain a homogeneous temperature distribution for the whole system, it avoids strange artefacts such as obtaining hot and cold parts of the molecule. When the heating is finished the production run could be started. In this run each of the UDP-GlcNAc conformers was simulated for about 100ns. This resulting production run trajectories were then analyzed to identify conformations of UDP-GlcNAc and transitions between them. The outlined protocol was applied for all of the 20 conformers in parallel leading 20 production run trajectories. The figure below shows a schematic representation of the simulation protocol.

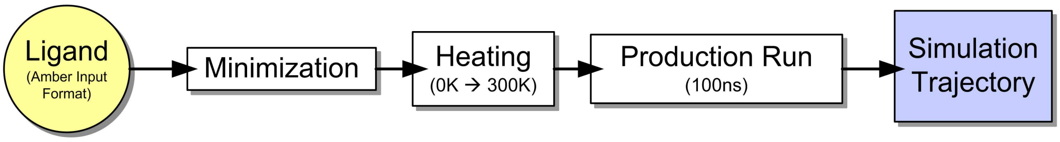


Figure 6: *In Vacuo* Simulation Workflow

3.2.2 Water

In the second simulation setup UDP-GlcNAc was simulated in a water environment to account for potential solvent effects. To model the water environment 3900 water molecules were placed around UDP-GlcNAc in an orthogonal box under periodic boundary conditions.

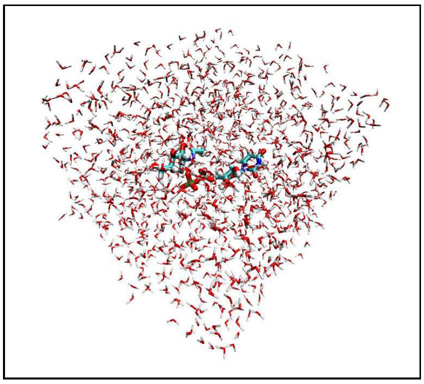


Figure 7: UDP-GlcNAc in Water Box

In comparison to the vacuum simulations, the computational complexity is largely increased in this scenario due to larger system size, resulting in a shorter total simulation time. The simulation protocol is the following. The initial minimization is a two stage process. First, only the water molecules are minimized and second the whole system is minimized. This procedure has been proven to prevent the simulations from steric solvent artefacts resulting from only minimizing the whole system. After minimization of the solvated system it was heated for 20ps from 0K to 300K. During the heating procedure the solute was weakly restrained to prevent it from wild fluctuations. To simulate the system as in a realistic experimental setting constant pressure conditions were turned on after the heating was completed. The system was then given 100ps to adapt its volume to the pressure constraint. When the pressure was equilibrated, production runs of 10ns were started. In order to reduce the computational complexity we used the triangulated TIP3 water model. In this model the H-O-H angle is fixed which reduces the calculation costs. Furthermore, the SHAKE [Ryckaert et al., 1977] algorithm was used to restrain the hydrogen bond fluctuations, which allows to increase the simulation step size from 1fs to 2fs. This simulation protocol was again applied to all 20 conformers leading 20 water simulation trajectories. The principal protocol is outlined below.

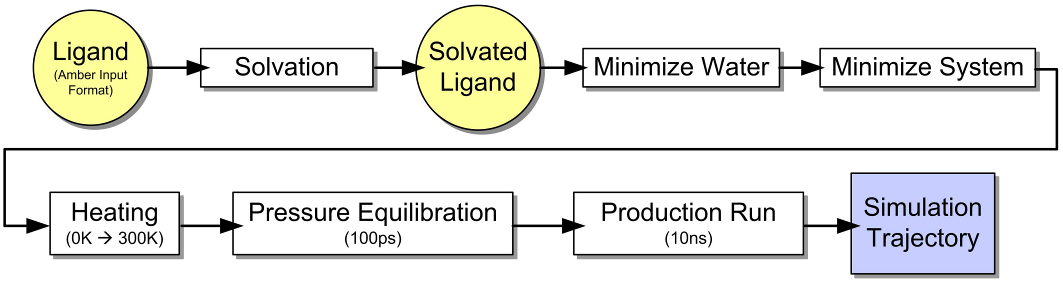


Figure 8: Water Simulation Workflow

3.2.3 Protein Environment

In addition to the vacuum and water simulations we also simulated UDP-GlcNAc in the binding site of the UDP-GlcNAc epimerase of E.coli. We decided to use the E.coli homologe as there was a protein crystal structure available from the PDB (PDB-Code: 1F6D)[Campbell et al., 2000], the structure of the human variant is currently only available as homology modeling based structure. The E.coli protein is present as a tetramer in the crystal data containing 4 identical units. For computational reasons we simulated only one subunit, assuming that this practice has no serious effect on the principal functionality. In the crystallographic data the UDP part of the ligand was also present, which allowed to determine the binding site of the enzyme. But as only information about the UDP part of the ligand was present, the GlcNAc orientation had to be modelled manually.

By assuming the UDP to be fixed at its position, the dihedrals 7,8 and 9 (refer to Figure 4) mainly determine the principal orientation of the GlcNAc moiety. Each of them were systematically turned in steps of 120° leading 27 different UDP-GlcNAc/UDP-GlcNAc-Epimerase complexes. For each of these complexes we performed short simulations in explicit water. The simulation protocol is similar to the ones used in the explicit water simulations. First, the surrounding water molecules where optimized, followed by the whole system. The minimization was followed by slowly heating up the solvated complex from 0K to 300K over 20ps, while putting weak restraints on the solute. When

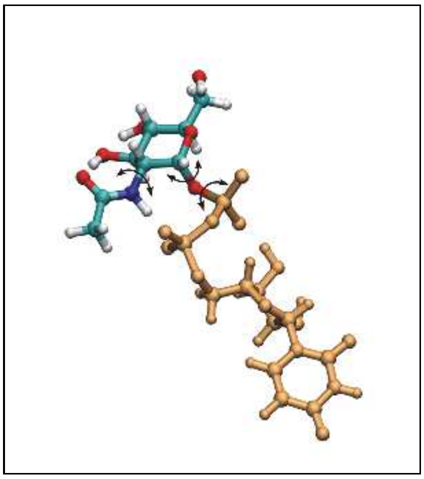


Figure 9: Dihedrals Turned to Model Potential Binding Conformations Based on Fixed UDP Part

the system was heated to 300K, pressure constraints were activated to obtain a pressure of 1 atm. The productive simulation run was started after giving the system 100ps to adjust to the pressure. Due to the severely increased system complexity (46646 atoms) computational restrictions allowed only a relatively short simulation of 100ps for each complex. Out of the 27 started simulations, 9 failed probably due to errors caused by energetic artefacts resulting from missoriented ligands.

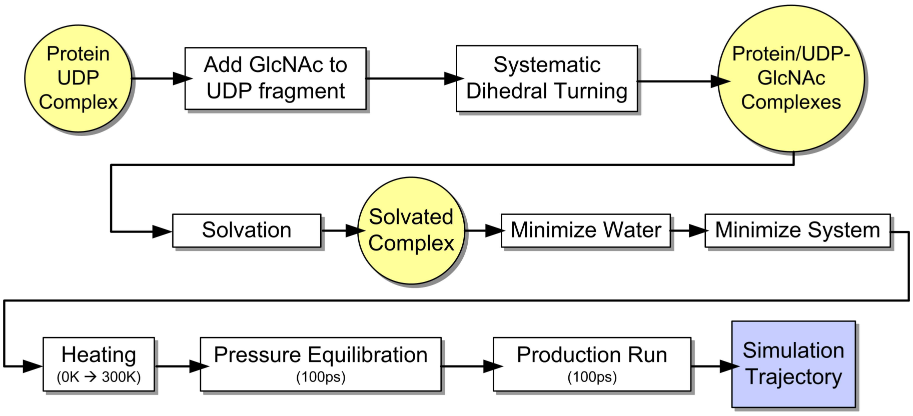


Figure 10: Protein Simulation Workflow

3.2.4 Summary

We performed in total 67 simulations in different environmental conditions using different simulation protocols. For an comprehensive overview of the Amber simulation input parameters refer to Appendix A. The simulations generated a number of trajectories containing the atom positions of UDP-GlcNAc over time. This data served as a basis for a comprehensive analysis.

	Atoms	Sim. Time	Step Size	Temperature	Pressure
In Vacuo	64	100 ns (2μs)	1 fs	300K	–
Water	3964	10 ns (200ns)	2 fs	300K	1 atm
Protein	46646	100 ps	2 fs	300K	1 atm

Table 3: Simulation Parameters

4 Data Analysis

Biomolecular systems and their dynamics can be characterized by the existence of biomolecular conformations. These conformations can be understood as metastable geometrical large scale structures that persit for long periods of time. On long time scales the dynamics of conformational changes can be regarded as flipping process [Elber and Karplus, 1987, Frauenfelder et al., 1989, Schütte et al., 1999, Schütte and Huisinga, 2003], whereas on shorter time scales a high flexibility in these conformations can be observed, resulting in a rich temporal multiscale structure [Nienhaus et al., 1992]. Biophysical research suggests that biomolecular systems often posses only a few dominant conformations, which can be described by only a small number of degrees of freedom [Amadei et al., 1993].

In many cases conformations can be characterized by analyzing the dihedral angles of a biomolecular system. In the present case we based our analysis on the dihedral angles of UDP-GlcNAc depicted in Figure 4. To obtain them, the generated MD trajectory data given in carthesian coordinates were projected into dihedral space. For the identification of metastable sets (conformations) from that data we mainly applied a hiddem Markov model (HMM) based approach [**?**] as well as Perron Cluster Cluster analysis [**?**]. Apart from the identification of metastable sets in the vacuum and water trajectories we also attempt to gain insights about the binding conformation of UDP-GlcNAc. Below the utilized methods are described in more detail.

4.1 HMM with Gaussian Output

HMMs have been used in a variety of application areas ranging from speech recognition over signal processing to Bioinformatics. More recently they have also been employed to identify biomolecular conformations based on dihedral angle time series data [**?**]. In that respects HMMs exhibit a remarkable feature. The analyzed essential coordinates not need to necessarily enable a geometric separation of the conformations, with HMMs also dynamical properties of the system are considered.

A HMM can be considered as a stochastic process with hidden and observable states (see Figure 11). The hidden process consits of a sequence of random variables $X_1, X_2, \ldots$ taking values from some state space, where X_t stands for the hidden state of the system at time t. *Hidden* state in this context refers to the fact that it cannot be observed, what is observed is just the output caused by a specific state. In the HMM setting it is assumed that the hidden process is a Markov process, i.e., the state X_n is memoryless and depends

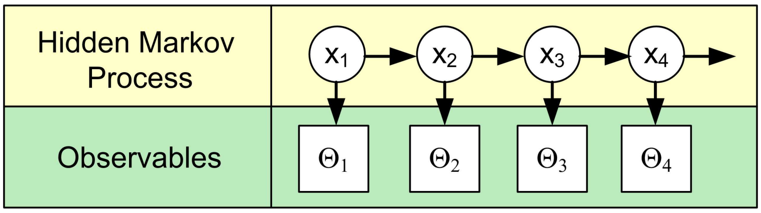


Figure 11: HMM Principle

only on its predecessor X_{n-1}. As the state space is assumed to be finite in the present application the Markov process can be characterized by a transition matrix $P = (p_{ij})$, where p_{ij} denotes the conditional probability of switching from state i to j. Note that P is a row stochastic matrix, meaning that each row adds up to one. Each hidden states X_1 has a specific output distribution which can be discrete or continuous. Hence, realizations of HMMs are concerned with two sequences, a sequence of observations and a sequence of hidden states. In the context of molecular dynamics data, the hidden states correspond to the metastable subset (conformation) the molecular system is in at a certain time, while the observations are the dihedral positions at that time.

To completely describe a HMM the number of hidden states (conformations), the transition matrix P, an initial distribution and the output distributions for each hidden state must be qualified. A way to obtain all these data is to make use of the well known *Maximum Likelihood Principle*, a standard technique for estimating parameters of a statistical model with respect to observed data. In setting considered here it can be explained as follows. Let λ denote the set of all parameters necessary to describe an HMM and $\Theta = \Theta_1, \Theta_2, \ldots, \Theta_T$ an observed sequence of data, i.e., in our setting dihedral values of UDP-GlcNAc. Let further denote $p(\Theta|\lambda)$ the probability[2] of observing the data sequence given the model λ. The likelihood function is then defined as $L(\lambda) = p(\Theta|\lambda)$, i.e., the observation is considered as being given and the functions asks for the variation of the probability in terms of the model parameters. The maximum likelihood principle states that the optimal parameters are given by the absolute maximum of L. Hence, finding the HMM that best describes the observed data can also be regarded as an optimization problem in the parameter space. In the context of HMMs this problem can be tackled by making use of an expectation-maximization algorithm (EM). Typically the determination of the optimal HMM parameters and the associated indentificaton of conformations involves three problems:

[2] strictly speaking $p(\Theta|\lambda)$ is actually a probability density function

I Calculation of the probability $p(\Theta|\lambda)$ for an observation sequence Θ (dihedrals) for a given λ (HMM).

II Estimation of the best model parameters for a given observation sequence.

III Given the estimated model λ and an observation sequence, determine the most probable hidden state sequence $X^* = (x_1^*, x_2^*, \ldots, x_T^*)$.

Problem I

The most straightforward method to calculate the probability $p(\Theta|\lambda)$ of an observation sequence given a model is to compute every possible hidden state sequence $X = (x_1, x_2, \ldots, x_T)$ of length T and sum up all their probabilities conditioned on the hidden sequence:

$$p(\Theta|\lambda) = \sum_{X=(x_1,\ldots,x_T)} p(X|\lambda)p(\Theta|X,\lambda)$$

Using this method would involve $2TN^T$ calculations which is infeasible for any practical application. By making use of so-called forward and backward probabilities it is possible to reduce this effort to TN^2. The method recursively divides the observation sequence Θ in two subsequences: one from tome 1 to time t and the other one from $t+1$ up to T. The forward probabilities are given by

$$\alpha_t(i) = p(\Theta_1, \Theta_2, \ldots, \Theta_t, X_t = i|\lambda),$$

denoting the probability of the observation sequence up to time t together with the information that the system is in hidden state i, conditioned to the given model λ. The backward probabilities are given by

$$\beta_t(i) = p(\Theta_{t+1}, \Theta_{t+2}, \ldots, \Theta_T|X_t = i, \lambda),$$

denoting the probability of the observation sequence from time $t+1$ to T, in this case under the condition that the system is in hidden state i at time t and on the model λ. Both probabilities can be computed by using recursive formulas within TN^2 operations each. Once the forward and backward variables are computed, finally the probability of the observation sequence Θ can be computed as

$$p(\Theta|\lambda) = \sum_{i=1}^{N} \alpha_t(i)\beta_t(i).$$

The forward and backward probabilities will be used further to estimate the best model parameters.

Problem II

There is no way to analytically determine the globally best model parameters that optimally explain the observed data. But it is possible to estimate a $\hat{\lambda}$ that locally maximized the likelihood $L(\lambda) = p(\Theta|\lambda)$. In the context of HMMs this can be achieved by using the Baum-Welch algorithm [Baum et al., 1970, Baum, 1972], which belongs to the class of expectation maximization (EM) algorithms [A.P. Dempster and Rubin, 1977]. EM algorithms are learning algorithms that iteratively improve an initial parameter set and converge to a local maximum of the maximum likelihood function. They consists of two steps, the Expectation step and the Maximization step. Starting from some initial model $\lambda^{(0)}$ the model is iteratively refined by cycling through expectation and maximization steps:

Expectation-step In this step, probabilities for being in hidden state i in time step t, given the observation Θ_t at that time and the current model λ^k are calculated. This can be efficiently done by calculating the joint probabilities making use of the previously calculated forward and backward variables.

The probability to be in hidden states i at time t can be expressed as

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j)$$

where $\xi_t(i,j)$ is the joint probability of being in i at t and j in $t+1$.

$$\begin{aligned}
\xi_t(i,j) &= p(X_t = i, X_{t+1} = j|\Theta, \lambda^{(k)}) \\
&= \frac{p(X_t = i, X_{t+1} = j, \Theta, \lambda^{(k)})}{p(\Theta, \lambda^{(k)})} \\
&= \frac{p(X_t = i, X_{t+1} = j, \Theta|\lambda^{(k)})}{p(\Theta|\lambda^{(k)})} \\
&= \frac{\alpha_t(i) p_{ij} f_j(\Theta_{t+1}) \beta_{t+1}(j)}{p(\Theta|\lambda^{(k)})}
\end{aligned}$$

Maximization-step After having calculated the probabilities for each hidden state given the observations and the model, reestimation formulas are used to get an improved model $\lambda^{(k+1)}$ The initial distribution $\pi_i^{(k+1)}$ is reestimated by

$$\pi_i^{(k+1)} = \gamma_1(i).$$

The transition probabilities between hidden states are reestimated by

$$p_{ij}^{(k+1)} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$$

The probability density function for each hidden state f_i has to be reestimated too, this is done via their maximum likelihood estimators. The observations Θ_t used to determine f_i^{k+1} have to with the probability $\gamma_t(j)$ of the hidden state. In the present application we assume that the output of the hidden states follows a Gaussian distribution. The maximum likelihood estimators to determine this distribution are mean ($\hat{\mu}$)and variance ($\hat{\sigma}$) and are reestimated as follows:

$$\begin{aligned}
\hat{\mu_i^{(k+1)}} &= \frac{\sum_{t=1}^{T} \gamma_t^{(k)}(i)\Theta_t}{\sum_{t=1}^{T} \gamma_t^{(k)}} \\
\hat{\sigma_i^{(k+1)}} &= \frac{\sum_{t=1}^{T} \gamma_t^{(k)}(i)(\Theta_t - \hat{\mu}_i^{(k)})^2}{\sum_{t=1}^{T} \gamma_t^{(k)}}
\end{aligned}$$

The E- and M-steps are iteratively repeated until either a predefined maximum of iterations is reached or the improvement of the likelihood $L(\lambda)$ gets smaller than a certain limit. It has been shown by A.P. Dempster and Rubin [1977] that the estimated model does not get worse ($L(\lambda^{k+1}) \geq L(\lambda^{(k)})$) and therefore approaches always a local maximum.

Problem III

Once the HMM has been estimated it is possible to determine the most probable hidden sequence. In the setting of analyzing molecular dynamic data this hidden sequence corresponds to the classification of the simulation trajectory in metastable (conformations) sets. To calculate the most probable hidden sequence the Viterbi [Viterbi, 1967] algorithm can be used. It computes for a given HMM model λ and an observation sequence Θ the most probable hidden path $X^* = (x_1^*, x_2^*, \ldots, x_T^*)$.

Let

$$\delta_t(i) = \max_{x_1,x_2,\ldots,x_{t-1}} P(x_1, x_2, \ldots, x_t = i, \Theta_1, \Theta_2, \ldots, \Theta_t, |\lambda)$$

denote the probability of being in hidden state i at time t after t observations. This quantity can be computed recursively by

$$\delta_t(i) = \max_{1\leq j\leq N}(\delta_{t-1}(j)p_{ji})f_i(\Theta_t)$$

When the probabilities for all states at all time steps has been calculated, it is straightforward to compute the most likely hidden state sequence. To do so an additional variable $\psi_t(i)$ is introduced. It containts the number of the argument that maximizes $\delta_t(i)$. The complete Viterbi algorithm is given below:

1) Initialization

$$\begin{aligned}
\delta_1(i) &= \pi_i f_i(\Theta_1), \quad 1 \leq i \leq N \\
\psi_1(i) &= 0
\end{aligned}$$

2) Recursion

$$\begin{aligned}
\delta_t(i) &= \max_{1\leq j\leq N}(\delta_{t-1}(j)p_{ji})f_i(\Theta_t) \\
\psi_t(i) &= \arg\max_{1\leq j\leq N}(\delta_{t-i}(j)p_{ji}) \\
& \quad 2 \leq t \leq T, 1 \leq i \leq N
\end{aligned}$$

3) Backtracking

$$\begin{aligned}
x_T^* &= \arg\max_{1\leq j\leq N}(\delta_T(j)) \\
x_t^* &= \psi_{t+1}(x_t + 1), \quad t = T-1, T-2, \ldots, 1
\end{aligned}$$

x_t^* can then be interpreted as the metastable set (conformation) the molecular system is in at time step t.

4.2 Identification of Metastable Sets

The inherent problem with HMMs is the *a priori* specification of the number of molecular conformations. In general, this number is not known in advance. To circumvent this problem, the HMM procedure can be started with a sufficiently large number of hidden states, greater then the expected number of conformations. Unfortunately, increasing the number of hidden states also increases the number of HMM parameters that have to be estimated by the EM algorithm, e.g., with d hidden states d^2 covariance parameters of the Gaussian distribution must be estimated. This parameter space increase causes a slower convergence behavior, which makes the pure HMM approach inapplicable for systems with many degrees of freedom. Hence, the HMM method is often used in a modified way, as described below. First, the high dimensional observation space is decomposed into low dimensional subspaces. For example if Θ contains all observed torsion angles, a possible decomposition would be to consider each single torsion angle, i.e., $\Theta^(j) = \theta_1^{(j)}, \theta_2^{(j)}, \ldots, \theta_T^{(j)}$ denoting the time series of the jth torsion angle. Second, each of this single dimensional time series is separately analyzed by the HMM method outline above, resulting in k Viterbi paths, where k is the number of torsion angles. Each paths represents the conformational dynamics as detected from information contained in the single time series $\Theta^{(j)} j = 1 \ldots k$. Third, these single paths are combined into a single *global* path with the following structure $x_t = (x_t^{(1)}, x_t^{(2)}, \ldots, x_t^{(k)})$. Then each state of the occurring Viterbi patterns gets unique state identifier assigned resulting in a one dimensional path s_t containing all the information from x_t.

The obtained global path can be regarded as a discretization of the state space which allows to deduce a reversible transition matrix $P = p_{ij}$ defined as

$$p_{ij} = \frac{\#(i,j) + \#(j,i)}{\#(i) + \#(j)}$$

where $\#(i,j)$ denotes the number of transitions from i to j and $\#(i)$ the number of states i in the path. Given this matrix, the aim is to define non-overlapping subsets such that transitions within these subsets are maximized and transitions between minimized. Each subset then corresponds to a long living metastable conformation. The local flexibility of the conformation is represented by frequent transitions between the sets, whereas conformational changes are represented by rare transitions between sets. The determination of these metastable sets and the number of sets can be achieved via inspection of the spectral properties of the transition matrix. Due to reversibility and row stochasticity , all eigenvalues of P are real and the metastable subsets can be identified by eigenvectors of eigenvalues close to the maximal dominant eigenvalue $\lambda = 1$. Meaning that the number of metastable sets is equal to the number of eigenvalues close to 1 (including one), while the rest of the spectrum is separated by a spectral gap from 1. To identify the corresponding subsets (among other methods) the sign structure of the eigenvectors of the identified eigenvalues can be exploited. In this case each sign pattern determines a unique subset. This type of metastability analysis is termed Perron Cluster

Cluster Analysis (PCCA). For a more comprehensive description see [?]. It is important to note, this way metastability is defined in respect to a certain lag time. This lag time arises from the frequency molecule coordinates are written to the trajectory file by the simulation program. When the lag time changes, the transition matrix inferred from the global Viterbi path might change as well resulting in a different metastability. In general, the lag time should be chosen such that the Markov property holds, i.e., a state in the global Viterbi path is not affected by the previous ones (memoryless).

4.3 Analysis Potential Binding Structures

To understand the enzymatic mechanism of the UDP-GlcNAc epimerase it would be advantageous to know the correct orientation of its ligand in the binding pocket. Unfortunately, only the position of the UDP part of the ligand in the binding pocket is known from 3D structural data [?]. To address this problem different possible orientations of the missing GlcNAc part have been modeled into the pocket and simulated for 100ps. The resulting trajectories have been analyzed in respect to the stability of the 3 dihedrals which contribute most to the yet structurally unsolved sugar arrangement.

4.4 Conformational Matching

The binding conformation of the UDP part of the ligand is known from 3D structural data of the UDP-GlcNAc-Epimerase (E.coli). To address the interesting question if this conformation is native, i.e., natively occurring without the influence of the protein environment, the water trajectory data was screened for the binding UDP structure. For this purpose the structural difference of the binding UDP part to the UDP part of UDP-GlcNAc structures was calculated for each frame of the water trajectory. Here, structural difference is expressed in terms of dihedral difference of the structures. A quantity measuring this difference is the dihedral residual minimal square distance (RMSD).
Dihedral RMSD for Two Structures

$$RMSD(\Theta^A, \Theta^B) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\min(360 - (|\theta_i^A - \theta_i^B|), |\theta_i^A - \theta_i^B|))^2}$$

This calculation resulted in a data set containing an RMSD value for each time step in the water trajectory, which was related to the identified conformations of UDP-GlcNAc, allowing to draw conclusions about a potential bindind conformations.

5 Results

5.1 Simulations in Different Environment

5.1.1 UDP-GlcNAc Vacuo Simulation

We simulated 20 conformations UDP-GlcNAc without any solvent present at 300K for $2\mu s$. Atom coordinates were written to the trajectory in intervals of 5ps. From the generated trajectory data 17 dihedral (as defined in Fig. 4) trajectories were extracted and served as basis for a metastability analysis. First, principal component analysis (PCA) [?] was applied to obtain the principal components of the input data. We selected the 6 first principal components contributing to 95% of the variance in the data. On each of this 6 principal components HMM-Gaussian was applied, assuming 4 hidden states per dimension. This resulted in 6 Viterbi paths which were combined into a single global Viterbi path containing 243 states. The spectrum of the corresponding transition matrix showed a spectral gap after 38 eigenvalues. Hence, PCCA clustering was applied to the transfer matrix to identify 38 metastabale sets. A large portion of these sets had a relative weighting below 5% causing us to to group these together using an intelligent clustering scheme (personal communication E. Meerbach). After the summarization, 5 metastable conformations remained. The accordingly colored dihedral trajectories are depicted in Figures 12,13. The identified conformations of UDP-GlcNAc are depicted in Table 4. It becomes apparent that the most compact structure has the greatest weight (66%), whereas the most spatially relaxed structure has the least relative weight (5.5%). This characteristics can be explained by energetically favored inner molecular hydrogen bonds which are present in the interfold structures. As further consequence, all identified conformations appear to be very stiff and show only a small local flexibility, transitions between the conformations are also very rare. Based on the metastable clustering of the trajectory, a transition matrix can be calculated that represents the transitions between the 5 conformations. The corresponding transition network is shown in Figure 14. Edges between the conformations indicate possible transitions. Due to the rare transition events, the calculated transition probabilities are extremely small, here in the range of 0.1%. As the transition matrix was build based on a time lag of $5ps$, the transition probability has to be also interpreted in that time scale.

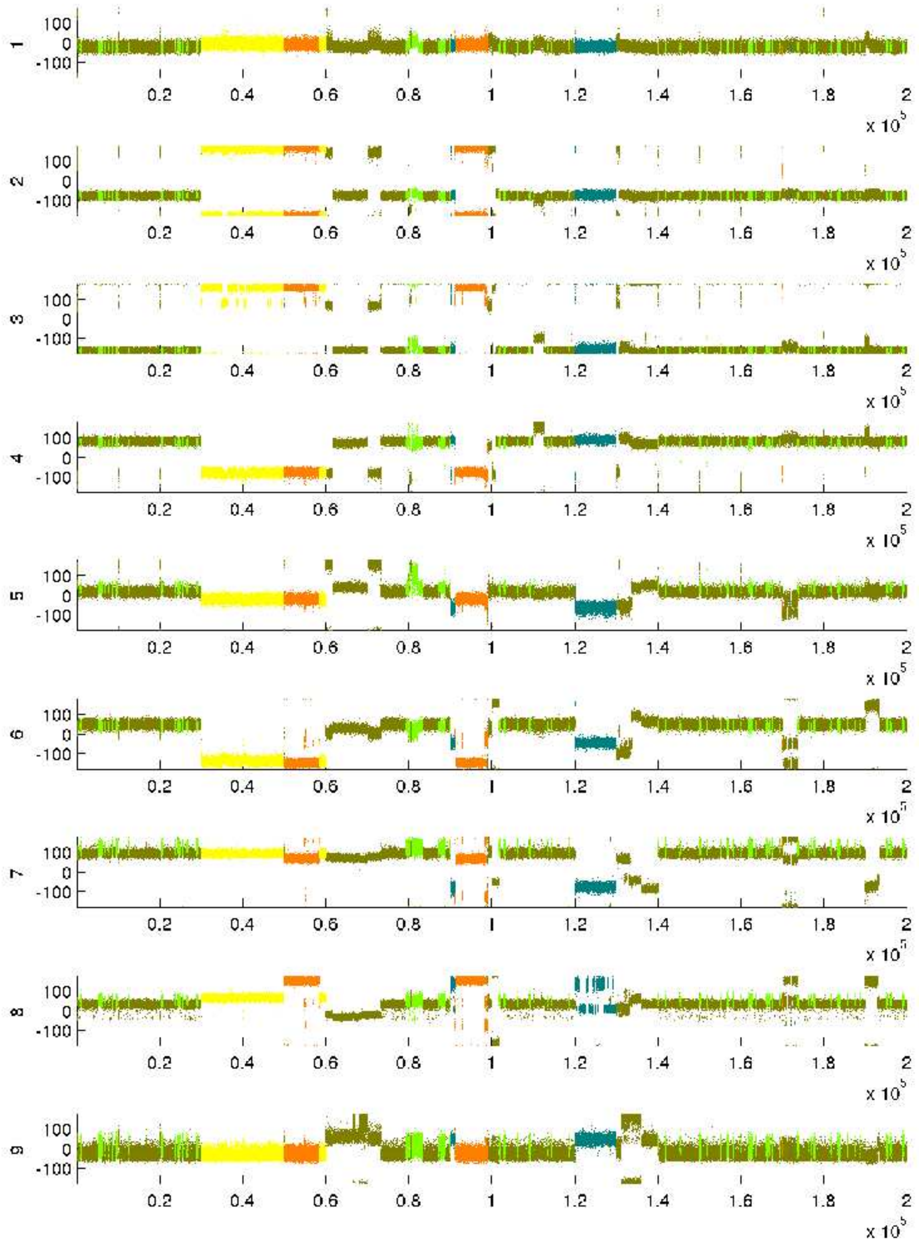


Figure 12: $2\mu s$ Vacuum Simulation Trajectory Data – Dihedrals 1-9

5.1.2 UDP-GlcNAc Water Simulation

We simulated 20 generated UDP-GlcNAc conformers in explicit water at a temperature of 300K and a pressure of 1atm. The simulations resulted in 20 trajectories of 10ns each. For the identification of metastable conformations all 20 trajectories were concatenated and analyzed using a similar methodology as for the vacuum data. Unfortunately, initial

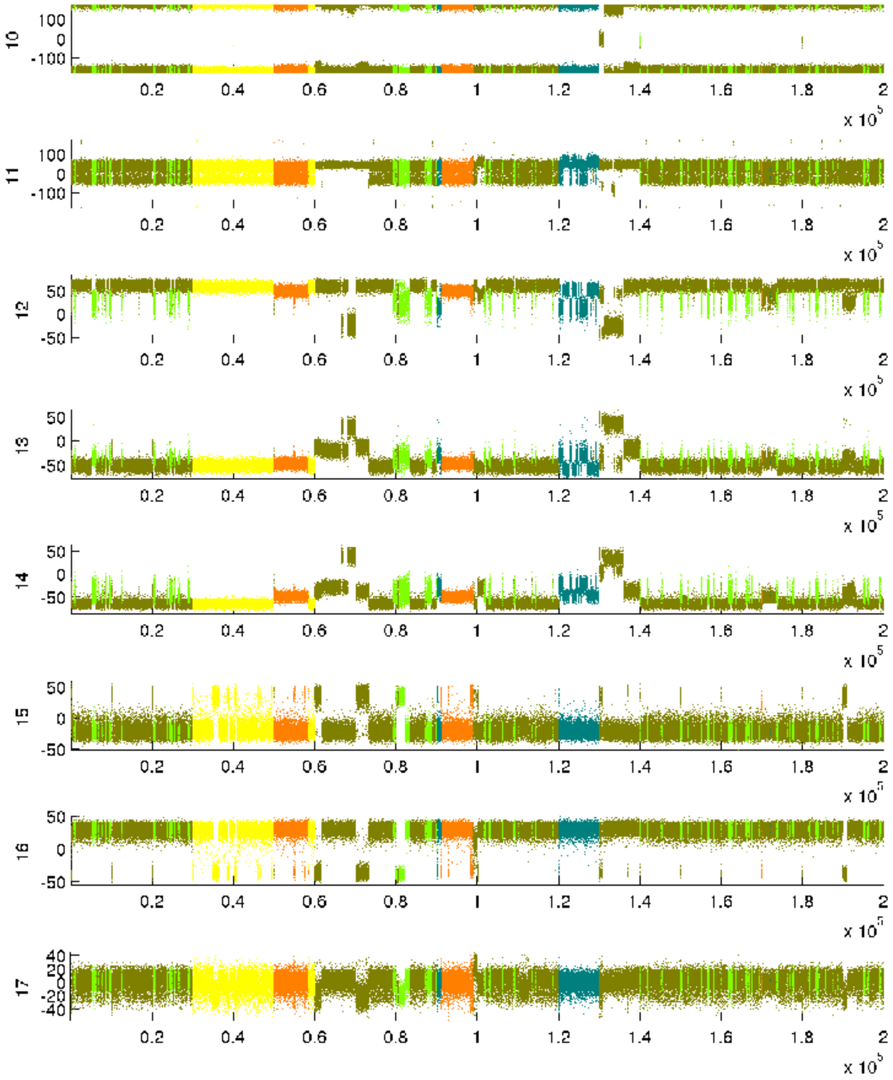


Figure 13: $2\mu s$ Vacuum Simulation Trajectory Data – Dihedrals 10-17

analyzes showed that the global conformations were dominated by the ring puckering of the GlcNAc and furanose ring making it difficult to obtain a meaningful clustering. Therefore, we choose only to consider dihedrals 1-8 for the analysis as they mainly determine the spatial properties of the molecule. On the trajectories of each of this 8 dihedrals HMM-Gaussian was applied assuming 3 hidden states. The resulting 8 Viterbi paths were combined into one, containing 1920 states out of $3^8 = 6561$ possibilities. To ensure the Markov property holds we calculated transition matrices[3] for different lag

[3] Transitions at the boundaries of concatenated trajectories were not counted in this case.

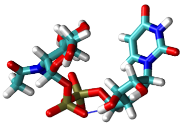
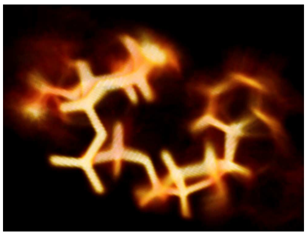
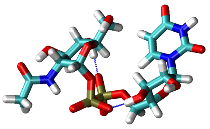
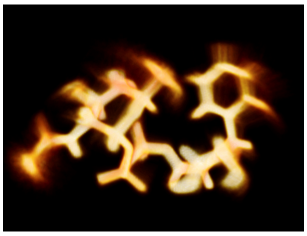
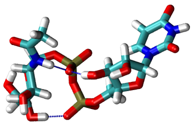
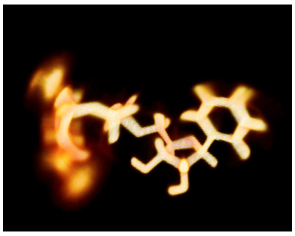
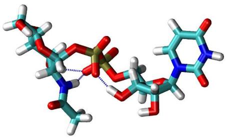
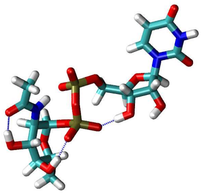
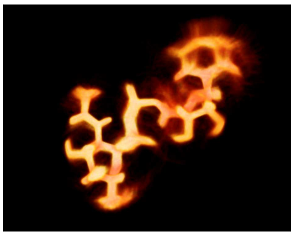
Conformation	Structure	Density Plot
C1 (66.6%)		
C2 (9%)		
C3 (8.1%)		
C4 (10.8%)		
C5 (5.5%)		

Table 4: Identified Vacuo Conformations with Relative Weighting

times and analyzed the corresponding spectra. The eigenvalue trend became constant at a lagtime of 160ps (80 steps), implying the observation is gets Markovian at this lag time. When looking at the spectrum of the corresponding transfer matrix a spectral gap after the 6th eigenvalue could be identified, suggesting 6 metastable sets. PCCA

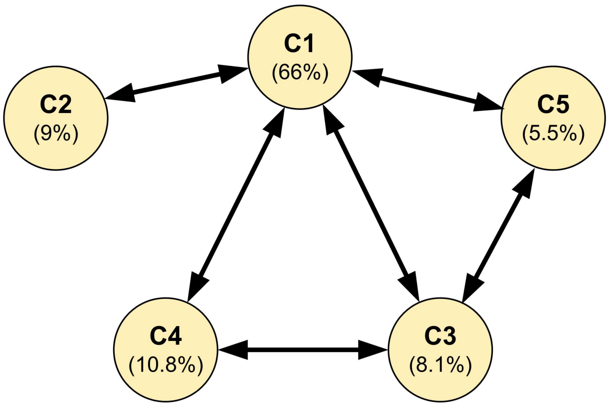


Figure 14: Transition Network Vacuo Conformations

was hence applied to the 1920x1920 transfer matrix to identify 6 metastable sets. The relative weightings of the identified conformations were $C1 = 58.3\%$, $C2 = 17.5\%$, $C3 = 11.9\%$, $C4 = 9.9\%$, $C5 = 1.5\%$ and $C6 = 1.0\%$.

They are depicted in Figure 17 which shows the transition network obtained from the clustered transfer matrix. In contrast to he conformations obtained from the vacuum simulations, here the most relaxed structure shows the highest relative weight whereas rather condensed structures have a small weight. Additionally, transitions between conformations are more frequent, e.g., conformation C6 changes to C1 with a probability of 9.1%, much higher than any transition probability in vacuum. The local flexibility of UDP-GlcNAc is also increased as it can be seen by the blurry parts of the density plots in the network graphic, making it hard to clearly identify structural features of the conformers.

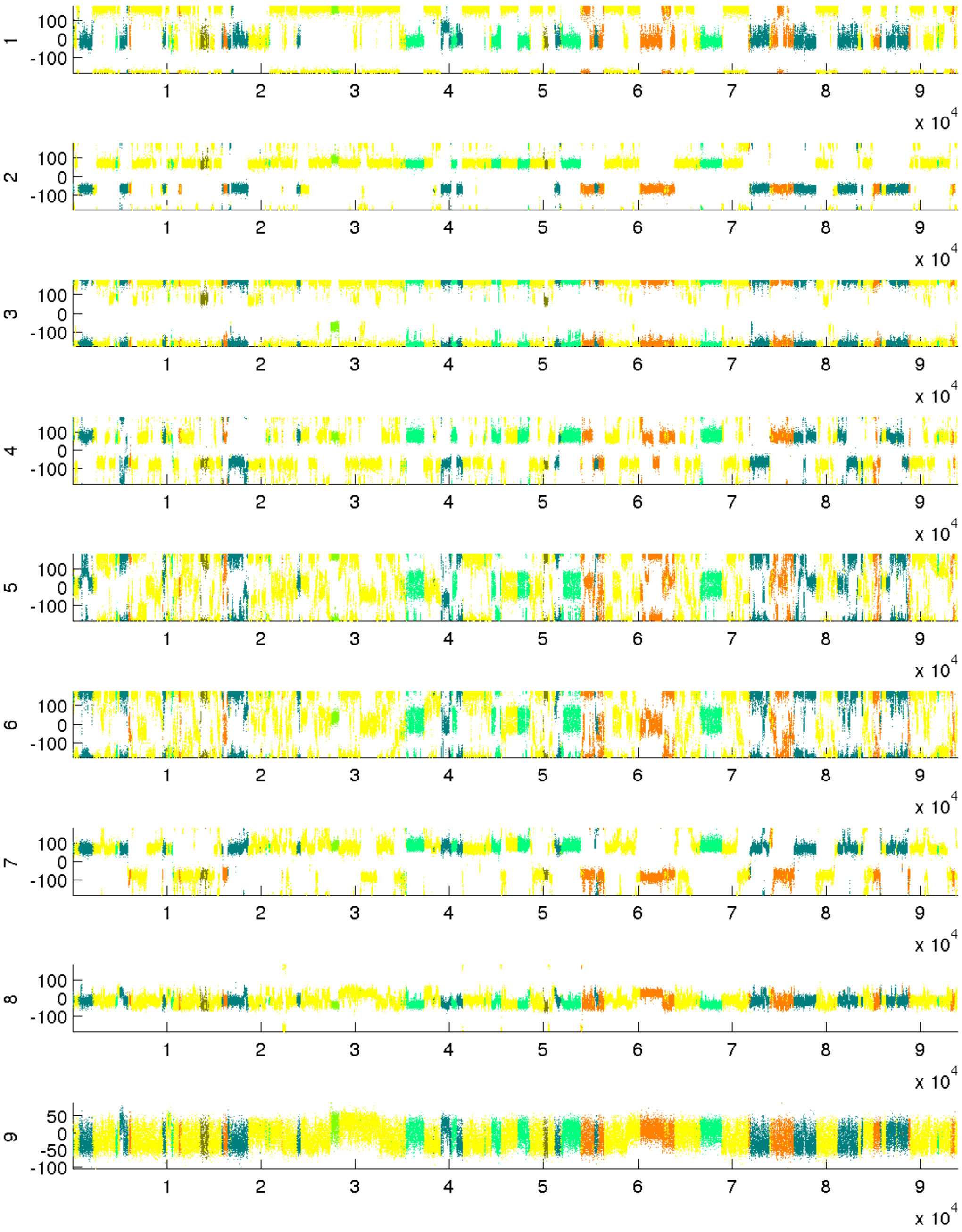


Figure 15: Water Simulation Trajectory Data – Dihedrals 1-9

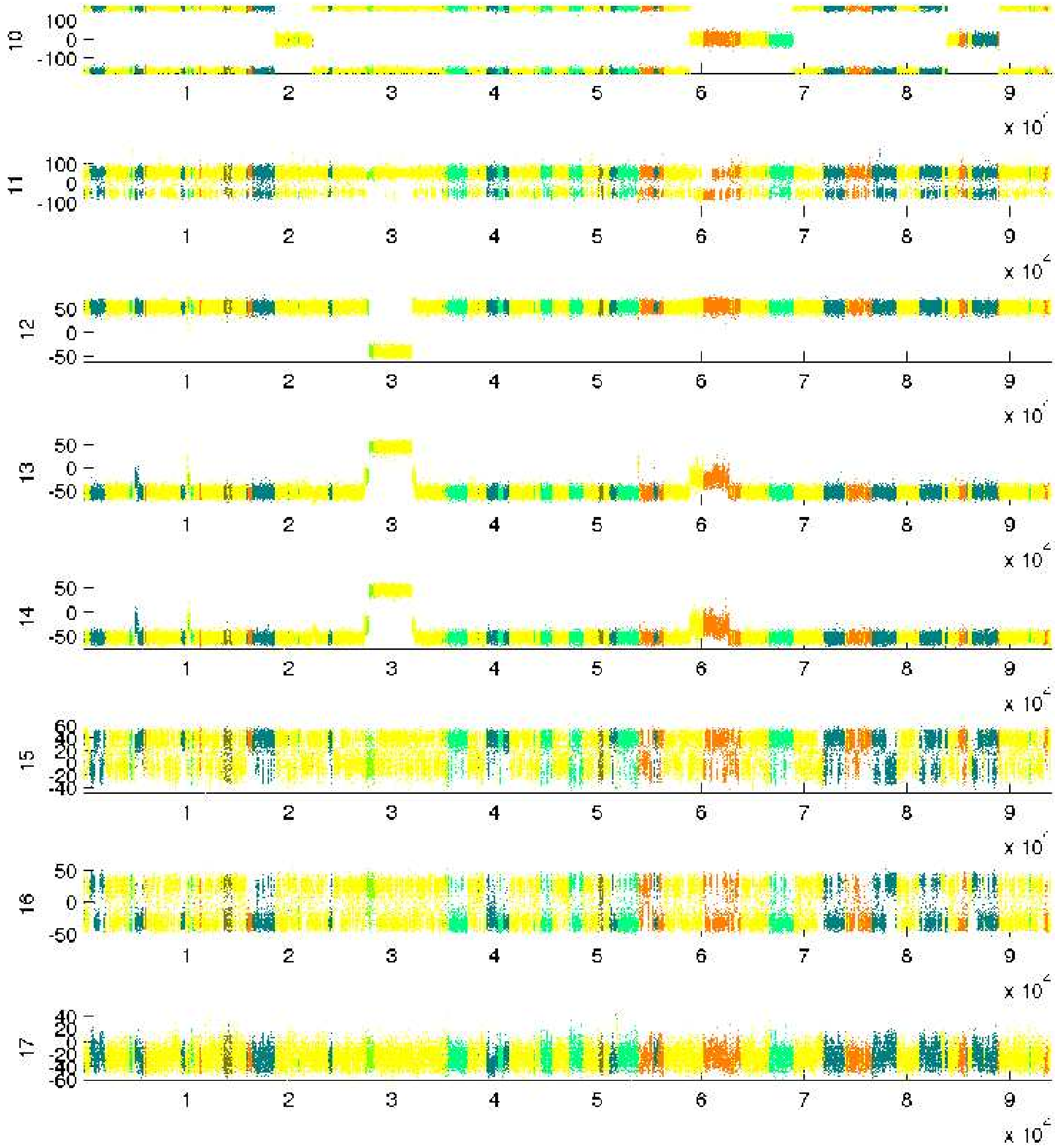


Figure 16: Water Simulation Trajectory Data – Dihedrals 10-17

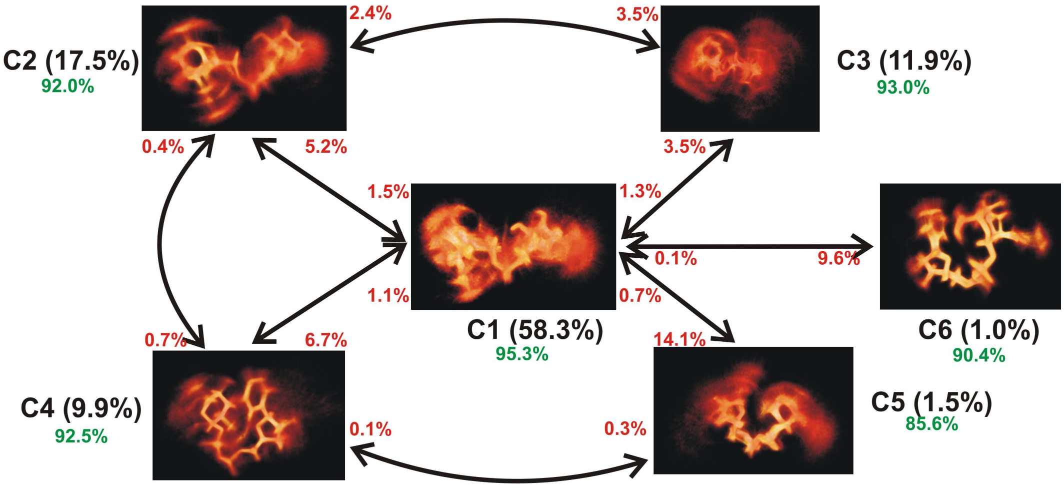


Figure 17: Transition network of UDP-GlcNAc in explicit water with density plots showing the flexibility within each metastable set. (in brackets – relative weighting, below brackets – persistence probability, next to pictures – exit probability) – Visualizations using Amira [?]

5.2 Discussion

The conformational behavior of UDP-GlcNAc shows clear differences depending on the environment, this can be directly seen from the dihedral trajectories (see Figures 12,13,15,16). Dihedral one, accountable for position the of the uracil ring show in vacuum almost no metastable behavior, in water, however, it it exhibits a high variability adopting two positions. Dihedrals 2-8, mainly responsible for the spatial extention of the molecule, don not show much variability in vacuum either. The different positions visible in the vacuum trajectories occur principally due to the different start conformations used. In the water environment these dihedrals exhibit a much stronger variability and metastability, which becomes apparent in the identified metastable conformations. They show a lot more local inner flexibility and changes between conformations are more frequent. The puckering of both rings shows also an interesting environment-dependent behavior. The GlcNAc ring (dihedrals 12-14) shows in water literally no puckering, whereas in vacuum it occasionally appears. In contrast, the furanose ring (dihedrals 15, 16) shows a great flexibility in water and is almost fixed in vacuum. It is furthermore remarkable that there is an anomalous change in dihedral 10 (peptide bond in N-Acetyl moiety), a behavior only observed in water. These fluctuations can probably be attributed either to a force field effect or to solvent interactions not present in vacuum. However, the flexibility of dihedral 11 (position of the -C6-O-H group) is coherent in both environments, only occasionally restricted in vacuum. Similar arguments apply for dihedral 17 (O-C1 bond in furanose ring).

Overall we observe ample conformations and high flexibility of UDP-GlcNAc in the water environment and few, stiff, well defined conformations in vacuum. This considerable difference can be attributed to solvent effects provoked by water molecules. In vacuum, UDP-GlcNAc forms several inner molecular hydrogen bonds stabilizing interfolded structures and reducing the flexibility. Such non-covalent bonds cannot develop in this way when water molecules are present, as they are rather formed with water molecules than with the molecule itself. Therefore, due to the absence of stabilizing bonds, conformations of UDP-GlcNAc are in water more relaxed and flexible.

The explanation for the complementary characteristics of the puckering of the ring systems is not as straightforward. A possible explanation for the puckering of furanose ring that only can be observed that strongly in water is given by the fact that frequent conformational changes in the "backbone" (dihedrals 2-8) directly affect the furanose arrangement. As there are not so many conformational changes in vacuum, the effect on the furanose puckering is limited in this case. The fact that there are no conformational changes observed for the GlcNAc moiety in water, is most likely an artefact of the short simulation times. The energy barrier of a conformational change in the sugar ring are too high to be overcome in this time. As the simulation time in vacuum is remarkably longer, observing a conformational change in the sugar is much more likely.

5.3 Testing Potential Binding Conformations

5.3.1 Protein - Ligand Stability

The simulation of 27 different binding conformations of UDP-GlcNAc in the binding pocket of the UDP-GlcNAc-Epimerase (E.coli) resulted in 16 100ps trajectories. 9 simulations failed due to errors caused by energetic artefacts resulting from missoriented ligands. The orientations of the three dihedrals (7, 8, 9) that determine the position of the GlcNAc moiety are displayed in Figure 18. For reasons of presentation all 16 trajectories have been concatenated, the alternating coloring corresponds to the different conformations. From the depicted data it becomes apparent that there are almost no significant changes in the dihedral angles of the respective conformations, the only apparent changes happen in dihedral 9 (orientation of the N-Acetyl Group) in conformations 5 and 12. Overall, this data indicates the existence of different binding conformations of UDP-GlcNAc.

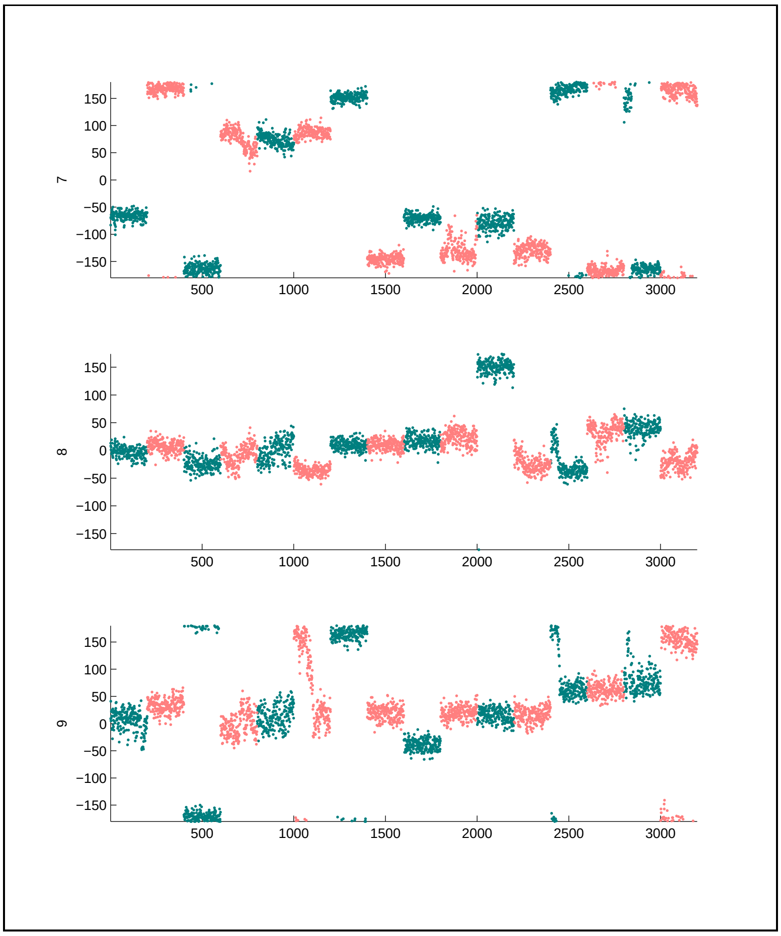


Figure 18: Concatenated simulation data of all protein environment simulations, only dihedrals that determine the orientation of the GlcNAc moiety are shown. The alternating colors correspond to the different simulated conformations.

5.3.2 Protein - Screening of UDP Conformation in Water

Simulating different conformations of UDP-GlcNAc in the binding pocket of UDP-GlcNAc-Epimerase did not reveal a favored binding conformation. To get nevertheless an idea of the most likely binding conformation we analyzed the water trajectory data to find the metastable set that contains the binding orientation of the UDP moiety. For this analysis we calculated the dihedral distance of the reference UDP (as it is bound to the protein) to all structures present in the water trajectory. The resulting distance over time plot is depicted in Figure 19. At certain times the dihedral distance gets rather small, in the range of 30 degrees. Hence, the corresponding structures can be seen as very similar to the reference UDP structure. Coloring the RMSD trajectory according to the determined metastable sets of the water trajectory it appears that a metastable conformation has associated most of the very small RMSD values.

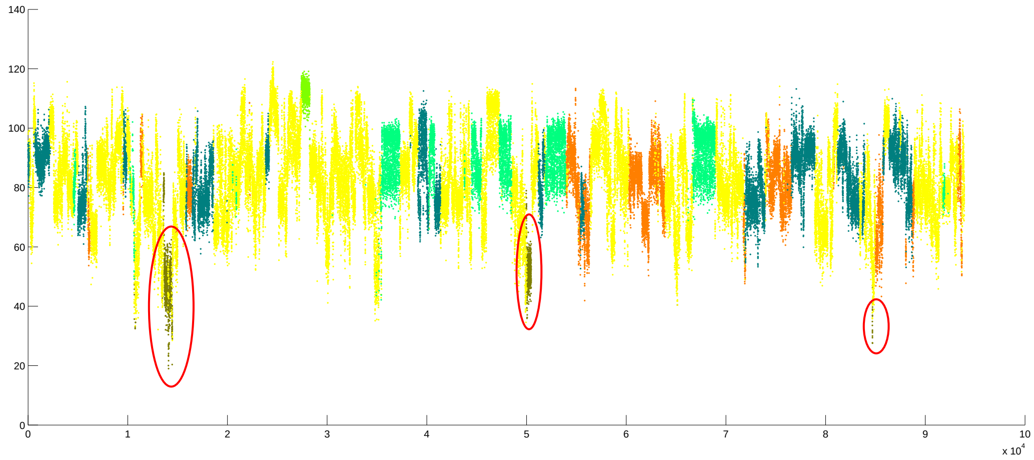


Figure 19: Dihedral RMSD of bound UDP to the UDP moiety in Water Displayed Over Time – Different Colors Correspond to Identified Water Conformations

More precisely it is the C5 conformation in respect to the conformational analysis of the water trajectory (compare to Figure 17). Thus leading to the conjecture that the C5 conformation might contain possible binding conformations of UDP-GlcNAc. In order to investigate this metastable set further, we extracted a sub-trajectory by taking all frames that correspond to the C5 subset. The UDP part in the sub-trajectory was then aligned to the reference UDP structure resulting in the density plot shown in Figure 20. The density plot shows a relatively stable UDP part, whereas the GlcNAc moiety shows strong local flexibility. Therefore, unfortunately, it is not possible to determine a favored GlcNAc orientation from the C5 metastable set and its density plot. But based on the statistical weight of the C5 conformation we can assume that at least the structure of the UDP part of UDP-GlcNAc occurs with a probability of 1.5% in the water environment.

5.4 Discussion

In order to rationally design an inhibitor for the UDP-GlcNAc epimerase it is of tremendous importance to understand the catalytic mechanism of the UDP-GlcNAc-2-Epimerase/ManNAc-Kinase. An inevitable step in that respect is determination of the correct binding position and orientation of the substrate. In case of the UDP-GlcNAc-2-Epimerase, the active site is known but the exact orientation of the substrate only partly in form of the UDP moiety of UDP-GlcNAc. This prompted us to model and test different possible

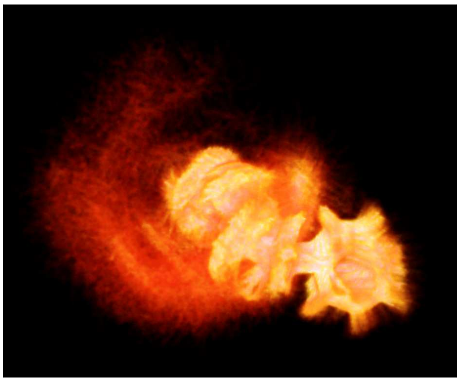


Figure 20: Density Plot of C5 Conformation Aligned to UDP Moiety

orientation of the GlcNAc moiety. The conducted simulations indicate that different orientations of GlcNAc are sterically possible. On the simulated time scale no transitions towards a particular structure is observed. From a purely geometrical point of view this therefore clearly indicates the possibility of different binding conformations. However, for the enzyme to be effective it is likely that solely a particular conformation is metabolized, which is most probable the energetically most favored one. Unfortunately, it is not possible to make reliable statements about the energy differences of the conformers because of high energy fluctuations and insufficient simulation times. In order to nevertheless gain insights about the correct binding conformations, the identified water conformations of UDP-GlcNac were related to the created potential binding conformations. Therefore, a geometric distance criteria was utilized to compare the structure of the bound UDP to the UDP structures in the water trajectory. At this analysis it turned out that several conformations (UDP moiety) in the time series are rather similar to the conformation of the bound UDP. Subsequent coloring of the trajectory according to the identified water conformations indicates that the majority of similar structures belongs to the C5 conformation of UDP-GlcNAc in water. If a well defined position for GlcNAc could be inferred from that conformation, it would be a clear indication for a likely binding conformation. To follow up on this question, all representatives of the C5 conformations have been aligned to the conformation of the bound UDP. Figure 20 shows the result, it becomes apparent that the UDP part of all representatives of C5 is quite similar but the GlcNAc moiety is not very well defined as implied by the blurry section. Based upon this result it is therefore difficult to make any statement about corresponding potential binding structures. An additional complete comparison (not restricted to the UDP moiety) of all conformations to all representatives of the C5 set revealed (data not shown) a few representatives with a small distance, however, the number was too small to draw any resilient conclusion. It is either possible that these were statistical outlier and the conformation does not exist in water, or the sampling of the conformational space is not sufficient to generate enough corresponding conformations. In case the binding conformation does not exist in "native" aqueous environment, it would be an indication for an *induced fit* binding mechanism, i.e., the enzyme has to alter the conformation of the ligand in order to bind it, which would be in contrast to a selective binding procedure, where the enzyme simply bind the appropriate conformation.

6 Conclusions

In this work we performed the first molecular dynamics study of UDP-GlcNAc, the key ligand of the sialic acid synthesis pathway, aiming at a better understanding of its conformational dynamics and determining its correct binding orientation when binding to UDP-GlcNAc-2-Epimerase (E.Coli). In order to be able to perform these studies the problem of missing molecule force field parameters was successfully solved by combining known parameters of two similar molecules. UDP-GlcNAc was simulated in vacuum, in explicit water and in the binding site of the enzyme. For an efficient sampling of the conformational space, different starting conformations were generated and corresponding simulations were distributed among several CPUs using a job scheduling system. By applying metastability analysis different conformations of UDP-GlcNAc could be identified from the simulation trajectories. Based on this data a considerable environmental effect becomes apparent. Conformations of UDP-GlcNAc without any solvent present differ substantially from conformations in aqueous solution. In water UDP-GlcNAc is observed to be much more flexible and unwound. Additional studies of potential binding conformations revealed that different orientations of the GlcNAc moiety of UDP-GlcNAc are possible in terms of their geometry. A subsequent analysis of the water trajectory unveiled the existence of the bound UDP conformation in the ”native” water environment. However, an unambiguous determination of the corresponding GlcNAc orientation was not yet possible. We hope that longer simulations will improve the sampling of the conformational space and therefore provide more insights into that question.

To sum up, the present conformational study could help in future to understand the initial step of the sialic acid pathway and therefore support the rational design of inhibitors.

7 Future Directions

Assuming it is possible to determine the binding rate of UDP-GlcNAc, this information could be used to further analyze the simulations for conformations having a lifetime in this range. The results might than provide additional insights about possible binding conformers and the binding process itself. Moreover, once the correct binding conformation is known and/or verified structural data for the human variant of the UDP-GlcNAc-2-Epimerase/ManNAc kinase is obtained, a quantum mechanical/molecular mechanics (QM/MM) simulation should be carried out. This type of simulations provide a powerfull tool to understand the actual reaction, as it is directly possible to simulate and observe actual enzymatic reactions, e.g., proton transfers. Furthermore, the effect of ions, e.g., Mg^{2+} or Mn^{2+} on the conformational behavior of UDP-GlcNAc should be investigated, recent work by Petrová et al. [2001] shows their importance in relation to UDP-Glc and glycosyltransferases. In general, longer simulations of the ligand in water environment appears to be desirable as the data suggests more sampling of the conformational space might reveal additional conformations which in turn might answer the question of the correct binding conformation.

Acknowledgements

First of all, I thank Christof Schütte and Werner Reutter for their scientific guidance and giving me the opportunity to work on such an exciting research topic. Stephan Hinderlich deserves thanks for the introduction into the world of sialic acids and his help with biochemical questions. Furthermore, I unconditionally thank the whole biocomputing group at the FU Berlin for their support and the terrific working atmosphere, especially Frank Noé and Eike Meerbach for continuously answering questions concerning molecular dynamic simulations and metastability analysis methods. A special gratefulness goes also to members of the Konrad Zuse Zentrum (ZIB), particularly Frank Cordes, Holger Meyer, Johannes Schmidt-Ehrenberg and Marcus Weber for assisting in generating conformers of the simulated ligand and familiarizing me with the visualization software Amira.

Finally, I am deeply indebted to my parents for thoroughly supporting me during all my years of study.

List of Publications From This Work

Proceedings and Poster Presentation

Martin Held, Eike Meerbach, Stephan Hinderlich, Werner Reutter, Christof Schütte. *Conformational Studies of UDP-GlcNAc in Environments of Increasing Complexity.* Workshop: From Computational Biophysics to Systems Biology 2007 (CBSB07). Forschungszentrum Jülich

8 Appendix - Amber Parameters

8.1 Parameters Vacuum Run

Minimization

```
&cntrl
   imin = 1,         // perform minimization
   maxcyc = 2000,    // maximal number of minimization cycles
   ncyc   = 1000,    // switch to conjugate gradient after ncyc steps
   ntb    = 0,       // periodic boundary conditions switched off
   ntr    = 0,       // no position restraining
   cut    = 12       // electrostatic cutoff 12 Angstroem
```

Heating

```
&cntrl
   imin  = 0,    // no minimization
   ntb   = 0,    // periodic boundary conditions switched off
   igb   = 0,    // no implicit solvent
   cut   = 12,   // electrostatic cutoff 12 Angstrem
   tempi = 0.0,// final temperature
   temp0 = 300.0,// inital temperature
   ntt   = 3,    // use Langevin dynamics to regulate temperature
   gamma_ln = 1.0, // collision frequency in Langevin equation
   nstlim = 20000, // total number of simulation steps (20ps)
   dt = 0.001,   // time step 1 fs
   nscm = 100,   // remove rotational and translational centre of mass
     // all 500 steps
   ntpr = 100,  // print out energy information all 100 steps (100fs)
   ntwx = 100,  // print out atom coordinates all 100 steps (100fs)
   ntwr = 1000  // print out restart file all 1000 steps (1ps)
```

Production

```
&cntrl
   imin  = 0,    // no minimization
   ntb   = 0,    // periodic boundary conditions switched off
   igb   = 0,    // no implicit solvent
   cut   = 12,   // electrostatic cutoff 12 Angstrom
   tempi = 300.0,// final temperature
   temp0 = 300.0,// initial temperature
   ntt   = 3,    // use Langevin dynamics to regulate temperature
   gamma_ln = 0.1, // collision frequency in Langevin equation
   nstlim = 100000000, // total number of simulation steps (100ns)
   dt = 0.001,   // time step 1 fs
   nscm = 500,   // remove rotational and translational centre of mass
     // all 500 steps
   ntpr = 2000,  // print out energy information all 2000 steps (2ps)
   ntwx = 2000,  // print out atom coordinates all 2000 steps (2ps)
   ntwr = 10000  // print out restart file all 10000 steps (10ps)
```

8.2 Paramters Water Run

Minimization Water Only

```
&cntrl
   imin = 1,       // perform minimization
   maxcyc = 2000,  // maximal number of minimization cycles
   ncyc   = 1000,  // maximal number of minimization cycles
   ntb    = 1,     // periodic boundary conditions switched on (const V)
   ntr    = 1,     // use positional restraining
   cut    = 10     // electrostatic cutoff 10 Angstroem
  /
Hold Solute Fixed // put restraints on the solute to keep it fixed
500
RES 1 1
END
END
```

Minimization System

```
Minimize Entire System
&cntrl
   imin = 1,      // perform minimization
   maxcyc = 2500, // maximal number of minimization cycles
   ncyc   = 1000, // switch to conjugate gradient after ncyc steps
   ntb    = 1,    // periodic boundary conditions switched on (const V)
   ntr    = 0,    // no position restraining
   cut    = 10,   // electrostatic cutoff 10 Angstroem
```

Heating

```
&cntrl
   imin  = 0,  // no minimization
   ntx   = 1,  // read initial coordinates from file
   ntb   = 1,  // periodic boundary conditions switched on (const V)
   cut   = 10, // electrostatic cutoff 10 Angstroem
   ntr   = 1,  // use positional restraining
   ntc   = 2,  // use SHAKE to constrain bonds involving hydrogens
   ntf    = 2, // omit force interactions involving hydrogens
   tempi = 0.0, // inital temperature
   temp0 = 300.0, // target temperature
   ntt   = 3,   // use Langevin dynamics to regulate temperature
   gamma_ln = 1.0, // collision frequency in Langevin equation
   nstlim = 10000, // total number of simulation steps (20ps)
   dt = 0.002,     // time step 2fs
   ntpr = 100,     // print out energy information all 100 steps
   ntwx = 100,     // print out atom coordinates all 100 steps
   ntwr = 1000     // print out restart file all 1000 steps
/
Keep Solute Fixed // put weak restraints on the solute to keep it fixed
10.0
RES 1 1
END
END
```

Pressure Equilibration

```
&cntrl
   imin  = 0,  // no minimization
   irest = 1,   // continue calculation from given input file
   ntx   = 7,   // read positions, velocities and box information from file
   ntb   = 2,   // perodic boundary conditions (const P)
   pres0 = 1.0, // reference pressure (1 atm)
   ntp   = 1.0, // constant pressure dynamics (isotropic pos. scaling)
   taup  = 2.0, // pressure regulation time (2ps)
   cut   = 10,  // electrostatic cutoff 10 Angstroem
   ntr   = 0,   // no positional restraining
   ntc   = 2,   // use SHAKE to constrain bonds involving hydrogens
   ntf    = 2,  // omit force interactions involving hydrogens
   tempi = 300.0, // inital temperature
   temp0 = 300.0, // target temperature
   ntt   = 3,   // use Langevin dynamics to regulate temperature
   gamma_ln = 1.0,  // collision frequency in Langevin equation
   nstlim = 50000,  // total number of simulation steps (100ps)
   dt = 0.002,      // time step (2fs)
   ntpr = 100,      // print out energy information all 100 steps
   ntwx = 100,      // print out coordinates all 100 steps
   ntwr = 1000      // print out restart file all 1000 steps
```

Production Run

```
&cntrl
```

```
   imin  = 0,   // no minimization
   irest = 1,   // continue calculation from given input file
   ntx   = 7,   // read positions, velocities and box information from file
   ntb   = 2,   // perodic boundary conditions (const P)
   pres0 = 1.0, // reference pressure (1 atm)
   ntp   = 1.0, // constant pressure dynamics (isotropic pos. scaling)
   taup  = 2.0, // pressure regulation time (2ps)
   cut   = 10,  // electrostatic cutoff 10 Angstroem
   ntr   = 0,   // no positional restraining
   ntc   = 2,   // use SHAKE to constrain bonds involving hydrogens
   ntf    = 2,  // omit force interactions involving hydrogens
   tempi = 300.0, // inital temperature
   temp0 = 300.0, // target temperature
   ntt   = 3,   // use Langevin dynamics to regulate temperature
   gamma_ln = 1.0,  // collision frequency in Langevin equation
   nstlim = 5000000,  total number of simulation steps (10ns)
   dt = 0.002,  // time step (2fs)
   ntpr = 1000, // print out energy information all 1000 steps
   ntwx = 1000, // print out coordinates all 1000 steps
   ntwr = 10000 // print out restart file all 10000 steps
```

8.3 Paramters Protein Simulation

Minimization Water Only

```
&cntrl
   imin = 1,       // perform minimization
   maxcyc = 2000,  // maximal number of minimization cycles
   ncyc   = 1000,  // maximal number of minimization cycles
   ntb    = 1,     // periodic boundary conditions switched on (const V)
   ntr    = 1,     // use positional restraining
   cut    = 10     // electrostatic cutoff 10 Angstroem
  /
Hold Solute Fixed // put restraints on the solute to keep it fixed
500
RES 1 372
END
END
```

Minimization System

```
Minimize Entire System
&cntrl
   imin = 1,      // perform minimization
   maxcyc = 2000, // maximal number of minimization cycles
   ncyc   = 1000, // switch to conjugate gradient after ncyc steps
   ntb    = 1,    // periodic boundary conditions switched on (const V)
   ntr    = 0,    // no position restraining
   cut    = 10,   // electrostatic cutoff 10 Angstroem
```

Heating

```
&cntrl
   imin  = 0,  // no minimization
   ntx   = 1,  // read initial coordinates from file
   ntb   = 1,  // periodic boundary conditions switched on (const V)
   cut   = 10, // electrostatic cutoff 10 Angstroem
   ntr   = 1,  // use positional restraining
   ntc   = 2,  // use SHAKE to constrain bonds involving hydrogens
   ntf    = 2, // omit force interactions involving hydrogens
   tempi = 0.0, // inital temperature
   temp0 = 300.0, // target temperature
   ntt   = 3,   // use Langevin dynamics to regulate temperature
   gamma_ln = 1.0, // collision frequency in Langevin equation
   nstlim = 10000, // total number of simulation steps (20ps)
   dt = 0.002,     // time step 2fs
   ntpr = 100,     // print out energy information all 100 steps
   ntwx = 100,     // print out atom coordinates all 100 steps
   ntwr = 1000     // print out restart file all 1000 steps
/
Keep Solute Fixed // put weak restraints on the solute to keep it fixed
10.0
RES 1 372
END
END
```

Pressure Equilibration

```
&cntrl
   imin  = 0,  // no minimization
   irest = 1,   // continue calculation from given input file
   ntx   = 7,   // read positions, velocities and box information from file
   ntb   = 2,   // perodic boundary conditions (const P)
   pres0 = 1.0, // reference pressure (1 atm)
   ntp   = 1.0, // constant pressure dynamics (isotropic pos. scaling)
   taup  = 2.0, // pressure regulation time (2ps)
   cut   = 10,  // electrostatic cutoff 10 Angstroem
   ntr   = 0,   // no positional restraining
   ntc   = 2,   // use SHAKE to constrain bonds involving hydrogens
   ntf    = 2,  // omit force interactions involving hydrogens
   tempi = 300.0, // inital temperature
   temp0 = 300.0, // target temperature
   ntt   = 3,   // use Langevin dynamics to regulate temperature
   gamma_ln = 1.0,  // collision frequency in Langevin equation
   nstlim = 50000,  // total number of simulation steps (100ps)
   dt = 0.002,      // time step (2fs)
   ntpr = 100,      // print out energy information all 100 steps
   ntwx = 100,      // print out coordinates all 100 steps
   ntwr = 1000      // print out restart file all 1000 steps
```

Production Run

```
&cntrl
   imin  = 0,   // no minimization
   irest = 1,   // continue calculation from given input file
   ntx   = 7,   // read positions, velocities and box information from file
   ntb   = 2,   // perodic boundary conditions (const P)
   pres0 = 1.0, // reference pressure (1 atm)
   ntp   = 1.0, // constant pressure dynamics (isotropic pos. scaling)
   taup  = 2.0, // pressure regulation time (2ps)
   cut   = 10,  // electrostatic cutoff 10 Angstroem
   ntr   = 0,   // no positional restraining
   ntc   = 2,   // use SHAKE to constrain bonds involving hydrogens
   ntf    = 2,  // omit force interactions involving hydrogens
   tempi = 300.0, // inital temperature
   temp0 = 300.0, // target temperature
   ntt   = 3,   // use Langevin dynamics to regulate temperature
   gamma_ln = 1.0,  // collision frequency in Langevin equation
   nstlim = 50000,  total number of simulation steps (100ps)
   dt = 0.002,  // time step (2fs)
   ntpr = 250, // print out energy information all 250 steps
   ntwx = 250, // print out coordinates all 250 steps
   ntwr = 250  // print out restart file all 250 steps
```

References

B. J. Alder and T.E. Wainwright. Phase transition for a hard sphere system. *J. Chem. Phys.*, 27(1208), 1957.

A. Amadei, A. Linssen, and H. Berendsen. Essential dynamics on proteins. *Proteins*, 17:412 – 425, 1993.

N.M. Laird A.P. Dempster and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. Roy. Stat. Soc.*, 39:1–38, 1977.

L.E. Baum. An inequality and associated maximiyation technique in statistical estimation for probabilistic functions of markov processes. *Inequalities*, 3:1–8, 1972.

L.E. Baum, T.Petrie, G. Soules, and N. Weiss. A maximization technique occuring in the statistical analysis of probabilistic functions of markov chains. *Ann. Math. Stat.*, 41:164–171, 1970.

H.J.C. Berendsen, J.P.M. Postma, W.F. Vangunsteren, A. Dinola, and J.R. Haak. Molecular-dynamics with coupling to an external bath. *J. Chem. Phys.*, 81:3684–3690, 1984.

V. P. Bhavanandan. Cancer-associated mucins and mucin-type glycoproteins. *Glycobiology*, 1(5):493–503, Nov 1991.

G. Blix. Über die Kohlenhydratgruppen des Submaxillarismucins. *Hoppe-Seylers Zeitschrift für physiologische Chemie*, 240:43–54, 1936.

R. S. Bresalier, R. W. Rockwell, R. Dahiya, Q. Y. Duh, and Y. S. Kim. Cell surface sialoprotein alterations in metastatic murine colon cancer cell lines selected in an animal model for colon cancer metastasis. *Cancer Res*, 50(4):1299–1307, Feb 1990.

R.E. Campbell, S.C. Mosimann, M.E. Tanner, and N.C. Strynadka. The structure of udp-n-acetylglucosamine-2-epimerase reveals homology to phosphoglycosyltransferases. *Biochemistry*, 29:14993–15001, 2000.

D. A. Case, T.A. Darden, T.E. Cheatham, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, K.M. Merz, B. Wang, D.A. Pearlman, M. Crowley, S. Brozell, V. Tsui, H. Gohlke, J. Mongan, V. Hornak, G. Cui, P. Beroza, C. Schafmeister, J.W. Caldwell, W.S. Ross, and P.A. Kollman. Amber 8. *University of California, San Francisco*, 2004.

W. Colli. Trans-sialidase: a unique enzyme activity discovered in the protozoan trypanosoma cruzi. *FASEB J*, 7(13):1257–1264, Oct 1993.

J. W. Dennis and S. Lafert. Recognition of asparagine-linked oligosaccharides on murine tumor cells by natural killer cells. *Cancer Res*, 45(12 Pt 1):6034–6040, Dec 1985.

K. Effertz, S. Hinderlich, and W. Reutter. Selective loss of either the epimerase or kinase activity of udp-n-acetylglucosamine 2-epimerase/n-acetylmannosamine kinase due to site-directed mutagenesis based on sequence alignments. *J Biol Chem*, 274(40): 28771–28778, Oct 1999.

R. Elber and M. Karplus. Multiple conformational states of proteins. *Science*, 235: 318–321, 1987.

C. Fahr and R. Schauer. Detection of sialic acids and gangliosides with special reference to 9-o-acetylated species in basaliomas and normal human skin. *J Invest Dermatol*, 116(2):254–260, Feb 2001. URL `http://dx.doi.org/10.1046/j.1523-1747.2001.01237.x`.

M. Fogel, P. Altevogt, and V. Schirrmacher. Metastatic potential severely altered by changes in tumor cell adhesiveness and cell-surface sialylation. *J Exp Med*, 157(1): 371–376, Jan 1983.

H. Frauenfelder, P.J. Steinbach, and R.D. Young. Conformational relaxation in proteins. *Chem. Soc.*, 29A:145–150, 1989.

S. Hakomori. Aberrant glycosylation in tumors and tumor-associated carbohydrate antigens. *Adv Cancer Res*, 52:257–331, 1989.

S. Hinderlich, R. Stsche, R. Zeitler, and W. Reutter. A bifunctional enzyme catalyzes the first two steps in n-acetylneuraminic acid biosynthesis of rat liver. purification and characterization of udp-n-acetylglucosamine 2-epimerase/n-acetylmannosamine kinase. *J Biol Chem*, 272(39):24313–24318, Sep 1997.

T. Ito, J. N. Couceiro, S. Kelm, L. G. Baum, S. Krauss, M. R. Castrucci, I. Donatelli, H. Kida, J. C. Paulson, R. G. Webster, and Y. Kawaoka. Molecular basis for the generation in pigs of influenza A viruses with pandemic potential. *J Virol*, 72(9): 7367–7373, Sep 1998.

T. Kageshita, S. Hirai, T. Kimura, N. Hanai, S. Ohta, and T. Ono. Association between sialyl lewis(a) expression and tumor progression in melanoma. *Cancer Res*, 55(8): 1748–1751, Apr 1995.

K. A. Karlsson. Microbial recognition of target-cell glycoconjugates. *Curr Opin Struct Biol*, 5(5):622–635, Oct 1995.

E. Klenk. Über die Natur der Phosphatide und anderer Lipide des Gehirns und der Leber in der Niemann-Pickschen Krankheit. *Hoppe-Seylers Zeitschrift für physiologische Chemie*, 235:24–36, 1935.

L. A. Lasky. Selectin-carbohydrate interactions and the initiation of the inflammatory response. *Annu Rev Biochem*, 64:113–139, 1995. URL `http://dx.doi.org/10.1146/annurev.bi.64.070195.000553`.

J.A. McCammon, B.R. Gelin, and M. Karplus. Dynamics of folded proteins. *Nature*, 267:585–590, 1977.

Holger Meyer, Sebastian Moll, Frank Cordes, and Marcus Weber. Conflow - a new space-based application for complete conformational analysis of molecules. Technical Report 06-31, Konrad-Zuse-Zentrum für Informationtechnik Berlin (ZIB), 2006.

G.U. Nienhaus, J.R. Mourant, and H. Frauenfelder. Spectroscopic evidence for conformational relaxation in myoglobin. *PNAS*, 89:2902–2906, 1992.

J. C. Paulson, J. Weinstein, and A. Schauer. Tissue-specific expression of sialyltransferases. *J Biol Chem*, 264(19):10931–10934, Jul 1989.

P. Petrová, J. Koca, and A. Imberty. Molecular dynamics simulations of solvated udp-glucose in interaction with mg2+ cations. *Eur J Biochem*, 268(20):5365–5374, Oct 2001.

Pavla Petrová, Jaroslav Koa, and Anne Imberty. Potential energy hypersurfaces of nucleotide sugars: Ab initio calculations, force-field parametrization, and exploration of the flexibility. *J. Am. Chem. Soc.*, 121:5535 –5547, 1999.

Y. Pilatte, J. Bignon, and C. R. Lambr. Sialic acids as important molecules in the regulation of the immune system: pathophysiological implications of sialidases in immunity. *Glycobiology*, 3(3):201–218, Jun 1993.

J.-P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *J. Comp. Phys.*, 23:327–341, 1977.

R. Sawada, S. Tsuboi, and M. Fukuda. Differential e-selectin-dependent adhesion efficiency in sublines of a human colon cancer exhibiting distinct metastatic potentials. *J Biol Chem*, 269(2):1425–1431, Jan 1994.

R. Schauer. Sialic acids and their role as biological masks. *Trends in Biochemical Sciences*, 10:357–360, 1985.

Tamar Schlick. *Molecular Modeling and Simulation.* Springer, 2002.

C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard. A direct approach to conformational dynamics based on hybrid monte carlo. *J. Comput. Phys.*, 151:146–168, 1999.

C. Schütte and W. Huisinga. Biomolecular conformations can be identified as metastable sets of molecular dynamics. In P.G. Ciaret and J.-L. Lions, editors, *Handbook of Numerical Analysis X*, Special Volume Computational Chemistry, pages 699–744. 2003.

B. R. Seavey, E. A. Farr, W. M. Westler, and J. L. Markley. A relational database for sequence-specific protein nmr data. *J Biomol NMR*, 1(3):217–236, Sep 1991.

R. Stäsche, S. Hinderlich, C. Weise, K. Effertz, L. Lucka, P. Moormann, and W. Reutter. A bifunctional enzyme catalyzes the first two steps in n-acetylneuraminic acid biosynthesis of rat liver. molecular cloning and functional expression of udp-n-acetyl-glucosamine 2-epimerase/n-acetylmannosamine kinase. *J Biol Chem*, 272(39):24319–24324, Sep 1997.

F.H. Stillinger and A. Rahman. Improved simulation of liquid water by molecular dynamics. *J. Chem. Phys.*, 60:1545–1557, 1974.

S. Tomlinson, L. C. Pontes de Carvalho, F. Vandekerckhove, and V. Nussenzweig. Role of sialic acid in the resistance of trypanosoma cruzi trypomastigotes to complement. *J Immunol*, 153(7):3141–3147, Oct 1994.

A. Varki. Diversity in the sialic acids. *Glycobiology*, 2(1):25–40, Feb 1992.

A. Varki. Biological roles of oligosaccharides: all of the theories are correct. *Glycobiology*, 3(2):97–130, Apr 1993.

A.J. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. on Inform. Theory*, 13:260–269, 1967.

Woods Group. *Glycam Parameters - GLYCAM Web.* Complex Carbohydrate Research Center, The University of Georgia , Athens , GA, 2007. URL `http://www.glycam.com/gl_params.html`.