Simulations of protein thermodynamics and structures with the PRIME20 model

Dissertation

zur Erlangung des Doktorgrades der Naturwissenschaften
(Dr. rer. nat.)

der

Naturwissenschaftlichen Fakultät II
Chemie, Physik und Mathematik

der Martin-Luther-Universität Halle-Wittenberg

vorgelegt von
Herrn M. Sc. Arne Böker
geb. am 30.11.1990 in Bielefeld

Halle (Saale), den 18.03.2019

Gutachter: Prof. Dr. Wolfgang Paul
Prof. Dr. Thomas Kiefhaber
Prof. Dr. Jutta Luettmer-Strathmann

Datum der öffentlichen Verteidigung: 04.11.2019

«The proletarians have nothing to lose but their chains.»

— K. Marx, recognising the biological importance of macromolecules [126]

Contents

Chapter 1

Introduction

1.1 Biochemical motivation

At the time of writing, the worldwide Protein Data Bank [1, 22, 23, 227] contains about 140 000 entries, coordinate files of proteins or protein segments whose folded structure has been identified. Despite this impressive number, the general problem of protein folding cannot be considered to be solved yet, as even influential protein scientists have to acknowledge [16, 49]. Ultimately, the aim is to understand and to be able to predict for a given sequence of amino acids, what configuration it is going to fold into and how it reaches this configuration from a denatured state – implying the assumption of a native state unambiguously determined by the primary structure [7].

In about 90 years since the idea of protein folding emerged [12], of course a substantial number of contributions towards answering this question have been published. Among these are the oft-cited papers written by Linus Pauling in 1951, in which he details the secondary structures of α-helix [156, 158] and β-sheet [157], which since then have been confirmed to be the most common elements of protein structure.

Going forward in a rough historical sketch, in 1968 Cyrus Levinthal famously asked how a protein is able to find its unique native state among a selection of about 10^{300} theoretically conceivable configurations within seconds and faster [112], a problem which was subsequently dubbed "Levinthal's paradox". As a solution, Levinthal suggested the existence of folding pathways [111], a concept which – at least in this sense – has largely been rendered obsolete by the model of a free energy funnel [15, 48, 51] in which the native state is either the global minimum or a sufficiently deep local free energy well. In the latter case, the global minimum may constitute an alternative, misfolded configuration, which is a possible cause for the formation of so-called amyloids.

The amyloid state is common to a large number of vastly different proteins [11, 60, 101]. It is a configuration in which multiple proteins aggregate to form large fibrils consisting of stacked β-sheets. Such a behaviour is desirable in some cases, for instance to store hormones in an unfunctional state [73] or to create durable substances like spider silk [8, 9], but in many other cases the loss of biological activity in the amyloid state, sometimes combined with a gain of toxic functionality in an unknown intermediate state, causes neurological disorders. Among the most notorious cases are Alzheimer's disease [76], Prion diseases like Creutzfeldt-Jakob or kuru [10, 151], and Huntington's chorea [225, 231].

Chorea Huntington, or Huntington's disease (HD), was first described by George Huntington [84] as a curious and singular hereditary case of chorea endemic to a community on Long Island. The term "chorea" for a "dancing disorder", i.e. a disease in which the afflicted person appears to be dancing uncontrollably, had been coined long before by Paracelsus [155] who identified various causes for the disorder (although medical knowledge has advanced since Paracelsus' time and therefore his interpretations do not always meet modern standards), but the hereditary nature of this specific form of chorea had

not been observed in medical science before. At the present day it is also known that the disease exists worldwide, not only on Long Island where Huntington made his observations.

About seventy years after Huntington's observation of a hereditary form of chorea, deoxyribonucleic acid (DNA) was identified as the carrier of hereditary, i.e. genetic information [14] and its structure was discovered another ten years later [220]. These events paved the way for a large interdisciplinary project called The Huntington's Disease Collaborative Research Group to identify the gene responsible for HD, 121 years after the first description of the disorder [193]. This gene, called IT15, contains a stretch of CAG trinucleotide repeats, the length of which is correlated with the occurrence, severity and age of onset of HD [57, 98, 182]. Among the group's findings is a threshold of 37 CAG units below which the disease never occurs, and above which its effects tend to be more pronounced for longer sequences.

The reason for this threshold appears to lie in the molecular structure of the protein expressed by the gene IT15, aptly named huntingtin. The trinucleotide CAG is a codon for the amino acid glutamine, hence the CAG repeat is translated into a polyglutamine (polyQ) sequence near the huntingtin N-terminus. This sequence appears to be largely responsible for a tendency of huntingtin fragments to aggregate in an amyloid fashion [214, 225, 231], and a recent medical treatment attacking and degrading the mutated huntingtin has been claimed by its creators to alleviate the disease progression [2, 186]. A similar aggregation procedure of polyQ segments in other proteins has been found to cause at least nine further neurological diseases [4, 161, 162, 201]. The length threshold is similar in most of these diseases, usually around 40 glutamine (or CAG) repeats, which strongly suggests a common mechanism.

Although it is clear that the mechanism involves aggregation of the polyQ sections, it has not been understood accurately yet. Wetzel et al. investigated the critical nucleus of aggregation of polyQ peptides. For Q_{23}, they observed a critical nucleus size of four chains, which decreases rapidly to a single chain in case of Q_{26} [93, 94, 225], suggesting a strong tendency of these only slightly longer chains to aggregate. Meanwhile, an interpretation of spectroscopy data by Crick et al. indicated no structure change at all around the disease threshold (which, as already noted, is clearly higher than 26 residues) [41], and Vitalis et al. found a theta temperature of 390 K for polyQ, indicating that even short such peptides are at least globular, if not even collapsed into an aggregatable configuration [201]. Diverging results like these show how unclear the mechanism of aggregation remains to this day, and a more detailed understanding of the structures formed by polyQ is required in order to fight HD or other polyQ diseases.

To add to the uncertainty about the structures of polyQ itself, neighbouring residues in the protein affect the structure formation as well, be it by keeping the polyQ segment out of a collapsed state, thus hindering its aggregation [25, 42, 106], or by being capable of aggregation themselves [94, 129, 225]. Furthermore, because polyglutamine alone is not water-soluble, experimental set-ups usually contain a number of polar flanking residues to enhance solubility [6, 41, 161, 201, 214], a necessity whose implications are not well documented: on the one hand these flanking residues obfuscate the view on the intrinsic behaviour of polyQ, on the other hand this intrinsic behaviour might not even be too relevant because the *in vivo* situation is affected by flanking residues too. In either case it might be instructive to compare the properties of pure polyQ to those of polyQ with flanking residues, which – due to the solubility issue – is only possible in simulation. Simulations of pure polyQ have been reported in various publications [67, 105, 124, 201, 202], but comparative simulations between polyQ and such experimental sequences are harder to come by.

These questions motivate the simulation of polyglutamine thermodynamics and structures, the results of which are collected in the present dissertation. The dependence of polyQ structures on the chain length is one research interest as well as the influence of flanking residues and further experimentally necessary additions. For the latter, a comparability to such experiments is desirable and has led to the choice to expand the polyQ sequence by two chromophores used in spectroscopy and by a five-residue solubility tail attached to the C-terminus of the peptides. Such a set-up has been and is being used by the lab of Thomas Kiefhaber [26, 63, 64, 135, 230] and has recently been applied to polyQ by Peter Enke and Michael Schleeger within this lab [62].

The CAG sequence in the IT15 gene, like in all genes related to polyQ diseases, is evolutionarily unstable. In an event during DNA replication called slippage, one DNA strand can bind to the other at a mistaken site, leading to the addition of nucleotides in the replica [85, 102, 110]. Repetitive nucleotide sequences like the CAG stretches facilitate slippage because they provide multiple matching binding sites, and are therefore easily elongated [86]. A length increase of the CAG stretches from generation to generation had already been observed by the HD research group [57, 182], but an attribution to slippage is only found in later literature [153, 159]. The process of evolutionary growth of CAG or polyQ tracts appears to be a reason for the very existence of polyglutamine diseases [231].

A somewhat similar event, which occurs during translation, is called frameshift. Like the DNA polymerase in replication, the ribosome can "slip" along the mRNA strand [45, 129], leading to the sequence of CAG triplets to be read as AGC or GCA instead (depending on the number of missed nucleotides), which are then translated to serines or alanines respectively. These effects have been shown to occur in HD and to modify the severity of the disease itself [45], making polyserine (polyS) and polyalanine (polyA) interesting targets of research in the same context.

Like polyglutamine, polyalanine tracts have been found to cause aggregation diseases [3, 5, 130]. This suggests an underlying general ability of repetitive amino acid sequences to aggregate, which links back to the idea of the amyloid state being thermodynamically more stable or at least comparable to a collection of peptides in their native states. No such diseases are known for polyserine, but the aggregation behaviour of polyS is argued to be even stronger, which might cause a higher cell toxicity than in the documented polyA or polyQ diseases, thus allowing only these "mildly" aggregating sequences in the genome [129].

Overall, research regarding polyserine is sparse, although it occurs relatively frequently in comparison to other repetitive amino acid sequences [85]. Various authors report polyS tracts to be disordered [27, 77, 82, 104] and mostly just acting as linkers between functional segments [82, 85, 190], although disordered regions like these can serve essential functions as well [18, 56, 58]. Single serine residues have been shown to destabilise α-helices [118, 147].

In contrast to polyserine, polyalanine sequences have been investigated extensively by many authors. They can either aggregate in a β-sheet conformation [3, 130, 183] or fold into α-helices [34, 83, 100, 116, 125, 172, 176, 185], depending on molecular context. The notorious helix structure has been reiterated in many simulations as well [21, 32, 146, 149, 208], but also transitions into β-fibrils have been investigated [50, 141]. Thus, polyA and polyS provide diverse reasons to be simulated: understanding the structures of polyS might be helpful in the context of polyQ disorders as well as for general protein folding research while polyA mostly serves to benchmark the model and simulation method, but its behaviour in the aforementioned experimental set-up might be interesting to see as well.

1.2 Computer physical motivation

In order to perform a simulation, the first requirement is a model which will be able to resolve and represent the desired properties of the system in question. Most protein models can be sorted into one of three classes depending on their level of detail. The highest level of detail is found in atomistic force fields, representations in which every atom (usually including solvent atoms) is considered individually. These models offer a high accuracy at the expense of large computational effort, which is not always available or appropriate. Despite their assumed accuracy, these models must be handled with care because different force fields can yield surprisingly variable results [33, 67, 71, 185].

Computation time is often a limiting factor, especially if properties beyond microsecond time scales are to be simulated. Such simulations require a simpler representation of proteins. Coarse-grained models, in which each amino acid is approximated as a single bead and the solvent is usually taken into account as a mean-field effect, have been applied successfully in many variations in the past [79, 170, 191, 197, 222]. Of course, the results depend strongly on the choice of model, even more so than

in atomistic models, and the level of detail is limited by the simple geometry. The rather complex layout of a polypeptide chain cannot be caught by a single bead per amino acid, which leads to an interest in models of intermediate resolution.

Intermediate-resolution models usually depict the backbone as three beads and coarsen the side chain into one [184, 187]. Many of these models are only capable of representing homopolymeric proteins, but in recent years, intermediate resolution models have emerged in which each side chain is assigned its individual interaction behaviour. Some of these models include the four-bead force fields PLUM [21, 175] and PRIME20 [36], the three-bead representations AWSEM [46] and AWSEM-IDP [226] or the variable three-to-eight-bead model PRIMO [72].

The apparent simplicity and elegance of PRIME20, whose authors managed to reduce the number of interaction energies to just 19 parameters from 171 possible pairs of interacting bead types, served as a motivation to apply it in the project whose results are described in this dissertation. Over the course of the project, the model turned out to be less elegant than expected, requiring several situational parameters, i.e. geometry adaptations which only apply to certain beads within the chain. The influence of these parameters on the model behaviour turned out to go beyond an aesthetic uneasiness and will therefore be discussed in an own chapter.

Another choice to be made in advance of a simulation project is that of a suitable simulation algorithm. PRIME20 has been and is being investigated using Discontinuous Molecular Dynamics, a variation of molecular dynamics specifically designed for discontinuous potentials [199, 204]. Where molecular dynamics methods try to sample configuration space in the same way as the real-world system, a more efficient method to do so is known as Monte Carlo sampling, based on a random, unphysical selection of configurations. Much like coarse-grained models in comparison to atomistic force fields, Monte Carlo simulations are faster than molecular dynamics, but the gain in speed comes at the expense of details, in this case at the expense of dynamics, making both types of algorithms suitable to answer different questions.

In the present project, a method called Stochastic Approximation Monte Carlo (SAMC) was used. SAMC promises to overcome the limitations of conventional Monte Carlo methods and to provide a complete picture of thermodynamics and configuration space. If the reasons for peptides to aggregate as amyloids are to be uncovered, this kind of completeness is crucial as the thermodynamic development of different populations – for example α-helix and β-sheet configurations – needs to be observed. Thus, the dissertation presents results on the thermodynamics and structure formation of the aforementioned peptides using the PRIME20 model together with the SAMC algorithm.

1.3 Structure of the dissertation

The remainder of the dissertation is structured as follows. The second chapter describes PRIME20 in detail before the third chapter treats the SAMC method and its implementation for this project. The discussion of results is divided into three chapters, of which the first (i.e. chapter 4) leads the way through the general evaluation methods available by virtue of SAMC to a general picture of the chain length dependence of the chain behaviour. Chapter 5 discusses the aforementioned details of the PRIME20 model, called “squeeze factors”, by creating and analysing variants of the model. The last results chapter refers to an experimental set-up and investigates the potential influence of spectroscopy dyes and a tail added to enhance solubility on structure formation and thermodynamics. The final chapters serve to summarise the findings and to provide an outlook to future projects.

Chapter 2

Model

A computer simulation generally consists of two elements: a model which – broadly speaking – represents a physical system as a set of numbers to be understood by the machine, and an algorithm which provides rules to the machine how to interact with these numbers. In a more physical formulation, the model may represent coordinates of objects, their sizes and interaction energies, and the algorithm may translate these energies into forces, provide a set of movement rules for these objects et cetera. In this work, a protein model named PRIME20 was used in combination with a simulation method called SAMC. The latter will be introduced in chapter 3 and the present chapter treats PRIME20, beginning at the real physical and chemical structure of proteins before giving an overview of protein models in general and detailing PRIME20. The chapter ends with a brief discussion of spectroscopy chromophores relevant to this work and their representation in PRIME20.

2.1 Proteins

Proteins constitute one of the main classes of biopolymers, defined as biological molecules consisting of several repeat units with a similar chemical structure (monomers). In the case of proteins, these monomers are the amino acids, linked by peptide bonds. "Proteins" are sometimes distinguished from "polypeptides" based on chain length and biological function, but since large proteins do not occur in this work and the basic physical concepts governing both are identical, the terms protein, polypeptide and peptide can and will be used interchangeably throughout the following chapters.

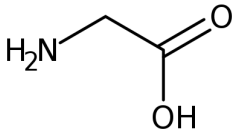

Figure 2.1 – Structural formula of glycine.

All so-called α amino acids[1] are made up of a common backbone and an individual side chain. The backbone consists of an amino (NH_2) group, the alpha carbon ($C_\alpha H_2$) and a carboxyl (COOH) group, as depicted in figure 2.1. This structure is equivalent to the simplest amino acid, glycine. The side chain substitutes one of the hydrogens bonded to the alpha carbon, as depicted in fig. 2.2 for the three amino acids which are most important for the present work. In an alanine molecule, the side chain is a methyl ($-CH_3$) group, in serine it is extended to hydroxymethyl ($-CH_2OH$) and in the case of glutamine it is a much longer chain with further substitutions. These three amino acids as well as Glycine (fig. 2.1) belong to the twenty proteinogenic amino acids which occur in natural proteins. Aside from these, there is a variety of synthetic amino acids, all identical in the backbone but distinguished by their individual side chains.

A protein is defined by its sequence of amino acids, called the primary structure. The spatial configuration it assumes is then categorised into secondary to quaternary structure, where secondary structure

[1] Other types of amino acids are not relevant in protein science, so the term "amino acids" always refers to α amino acids.

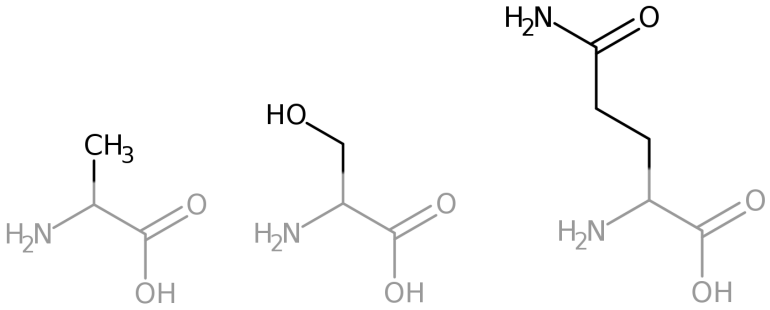

Figure 2.2 – Structural formulae of the amino acids alanine, serine and glutamine. The backbone, common to all amino acids and practically identical to Glycine (fig. 2.1), is coloured grey.

refers to basic recurring motifs which in their sum make up the complete three-dimensional tertiary structure. The term quaternary structure refers to the assembly of several protein units into a large complex. Even though the amyloid conformation, which motivates the scientific questions in the present work, could be regarded as a quaternary structure, the simulations carried out so far only employed single small peptides, hence quaternary structure does not occur. Furthermore, tertiary and secondary structure are largely synonymous as these short chains are mostly unable to form configurations with multiple structural domains.

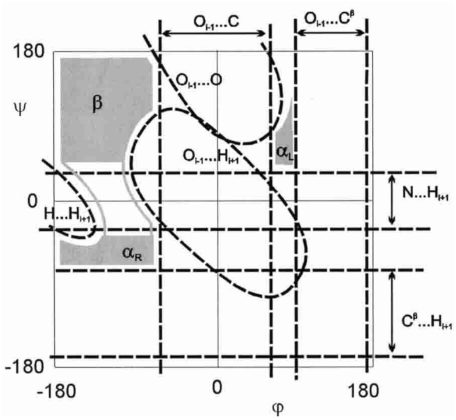

Figure 2.3 – Structure of a Ramachandran plot (Ho et al., 2003 [80, fig. 1]).

Secondary structure motifs are usually identified by Ramachandran plots or hydrogen bonds. The Ramachandran plot [80, 168] describes the backbone geometry by a set of dihedral angles of the three recurring bonds (C–N, N–C_α, C_α–C). The peptide bond (C–N) is always planar, so its dihedral angle ω can only take on values around 180° (also called *cis* configuration) or 0° (*trans*). However, the N–C_α and C_α–C dihedral angles, named ϕ and ψ respectively, are more variable. In a Ramachandran plot, each amino acid is depicted as one point on a (ϕ, ψ) map. Certain regions of this map correspond to different secondary structure motifs, as shown in the scheme in figure 2.3.

The second method of identification is based on backbone hydrogen bonds (H-Bonds) between the amide nitrogen and carboxy oxygen atoms. The most common motif, the α-helix, is characterized by regular H-Bonds between the $(i + 4)$-th amide hydrogen and the i-th carboxy oxygen for any i, while the other important motif, the β-hairpin, consists of two antiparallel strands connected by a β-turn. Such a turn is stabilised by a non-repeating $(i + 3, i)$ H-Bond; the antiparallel strands themselves are characterised by the H-Bonds $(i + 5, i - 2)$, $(i + 7, i - 4)$ etc. and their counterparts $(i - 2, i + 5)$ etc. The interjacent available groups point to the outside, leaving them available for H-Bonding to further β strands. Aside from α-helix and β-sheet/turn, less common secondary structures include the π helix with recurring $(i + 5, i)$ H-Bonds, the 3_{10} helix and the 2.2_7 helix/ribbon (following the nomenclature by Bragg, Kendrew and Perutz [31]) with $(i + 3, i)$ and $(i + 2, i)$ H-Bonds, and two types of polyproline helices which are not stabilised by H-Bonds. The absence of regular structure is called random coil.

2.2 Protein modelling

To simulate the behaviour of a protein, a model is required which provides a representation of the protein atoms' positions and of the various interactions between those atoms. Perhaps the most intuitive way to achieve this is called atomistic modelling. As the name indicates, every atom is represented individually as a single bead. Their interactions are usually simplifications of the underlying electronic behaviour. This method provides a high level of detail which can only be increased by explicitly considering the quantum mechanical behaviour of the electron shells in so-called *ab initio* simulations.

While atomistic modelling promises detailed results, it is also quite resource-intensive. Even small peptides contain three-digit numbers of atoms, and a high number of solvent molecules need to be included in the simulation as well. Continuously updating the positions and interactions of such a large number of objects requires considerable computational time, which places a limitation on the real time scales accessible by such a method. This has led to the concept of coarse graining, where several atoms are either mapped to one bead or to one point on a lattice. Both variants increase computational efficiency by reducing the number of individual objects and interactions at the expense of details which are assumed to be negligible.

The oldest and – again – most intuitive coarse-grained protein representations are one-bead models where each bead represents one amino acid. The level of detail in these models is minimal, but basic structural features can be reproduced and the simplicity of these models makes it possible to single out the influence of distinct parameters on the physical behaviour. For example, Magee et al. [119] used a one-bead homopolymer model to find generic helical structures depending on the interaction distance of the monomers, and Taylor et al. [191] reported a folding behaviour thermodynamically comparable to that of short peptides in a similar model.

While these homopolymer models are successful in modelling aspects of protein behaviour, they are extremely generic and can be applied to non-protein polymers as well. To achieve a better specificity for proteins, further details need to be included. One example of this is the HP model, first proposed by Dill [47], where every bead is either considered hydrophobic or polar and the bead interactions depend on this property. Because most proteinogenic amino acid side chains can be classified as either hydrophobic or polar, this model represents one of the vital features of proteins and has been used and adapted in various ways. Another option is the inclusion of bead-specific interactions [222, chapter 6], which provides a basic form of backbone hydrogen bonding.

Models like these are called "physics-based" because their interaction potentials are founded in physical properties. The opposite strategy is called "knowledge-based" and comprises models whose interactions are derived from the knowledge of secondary or tertiary structures. The basis of many knowledge-based models is the Gō model [29, 189]. In this model, pair interactions are assigned based on proximity of beads in a native protein configuration. These interactions are not derived from a physical concept, but the native state arises from physical forces which are assumed to be mostly identical to Gō interactions [70, 148]. Gō-like models are expectably successful in folding the chain into its native state and they tend to reproduce protein behaviour in and around this state well. It should be kept in mind however that these models by definition require a native state, making them less suitable for the simulation of disordered protein states.

As these few examples show, one-bead models are highly variable and adaptable to the problem at hand. Due to the low number of beads they also require much less computational power than atomistic models, however the atomistic detail, which may be crucial, is lost. Calculating the dihedral angles to produce a Ramachandran plot, for example, requires several backbone beads. Similarly, even though specific interactions can be included in one-bead models, the distinctive directionality of hydrogen bonds cannot be expressed by a simple spherical bead. Coupled with the general increase in computational power over the years [136], weaknesses like these gave rise to so-called intermediate-resolution models, consisting of a number of beads per amino acid between one and the total number of atoms.

A popular class of intermediate-resolution models divides every amino acid into four beads. The geometric basis of these models was first proposed by Sun [184] and adapted and refined in several later models since then [35, 50, 137, 187, 205]. The backbone of such a four-bead model consists of three "united atoms", representing the NH, $C_\alpha H$ and CO groups, and a fourth bead for the side chain. This geometry gives access to the ϕ and ψ angles and allows to introduce a more realistic hydrogen bond geometry (detailed in the following section). The side chain representation as a single bead is coarse-grained and keeps the computational cost low in comparison to the all-atom models.

Side chain interactions are often classified as hydrophobic and polar, similarly to the one-bead HP model [47]. However, attempts have been made to represent all proteinogenic amino acid side chains individually, resulting in the models PLUM [20] and PRIME20 [36]. Contrary to the physics-based HP models these can be described as knowledge-based since the side chain interaction parameters are derived from folded structures, but unlike the Gō model, which always refers to a single native state, they are averaged from larger data sets and thereby generalized so these models are thought to be applicable to all proteins.

The one-bead representation of the side chain shares similar drawbacks with the one-bead representation of the whole amino acid. The side chains can have very different sizes (see fig. 2.2 for example), which are not trivial to model and among the models named here only occur in PRIME20, and all side chain interactions including hydrogen bonds are averaged to a single, type-specific potential. If required, these issues can be resolved by a more complex side chain geometry as seen in some sources [72, 87, 92, 122], which again increases the computational cost. On the "philosophical" side it might be interesting that these more complex models are usually physics-based again. As Miyazawa and Jernigan [134] (whose energy parameters the creators of PLUM [19, 20, 21] use) note, this fact is more than a technicality because *«[...] even though these effective potentials have the important characteristics of low values for the native folds of proteins, they are unlikely to succeed in representing the actual potential surface far from the native conformation. Therefore, such potentials of mean force may not be appropriate for application in a study of a wide range of conformations, from the denatured state to the native conformation [...]»*. Even though they are not biased towards a single native state like the Gō model, knowledge-based models like PLUM and PRIME20 will most likely favour folded states over disordered ones and especially for disordered proteins the quality of simulation results is uncertain [226], which is an issue to be kept in mind.

2.3 PRIME and PRIME20

For the present project, the choice was made to apply the four-bead model PRIME20 [36]. As noted before, PRIME20 provides parameters for all proteinogenic amino acid side chains, promising a sensible representation of any individual primary structure. The model uses discontinuous potentials well suited to Monte Carlo simulation, and a physical grouping of side chains supposedly alleviates the bias towards native structures mentioned in the previous paragraph while also reducing the number of energy parameters to a minimum. Because their values were obtained by means of an optimisation algorithm, a low number of parameters is desirable to reduce the risk of "trapping" in a locally optimal parameter set.

PRIME20 extends the earlier four-bead model PRIME to include the twenty proteinogenic amino acids. Thus, individual bead positions, sizes and interactions of the side chains did not exist in PRIME yet, but the backbone geometry is identical in both models. Reflecting this development, the following two subsections will treat the geometry and interactions of PRIME and the extension to PRIME20 is addressed afterwards.

PRIME geometry

The geometry of PRIME is depicted in figure 2.4. Each sphere stands for one of the united atoms: blue for the amino group (NH), green for the alpha carbon (C_α), red for the carboxy group (CO) and grey for the side chain (R). In comparison to real proteins, the backbone bead positions correspond to the respective N and C atoms and the side chain bead is placed at the center of mass. Although they are depicted identically here, the beads have different sizes depending on their type. These diameters and all further parameters are collected in section 2.4. The beads are connected by bonds (white sticks). Instead of a bond angle potential, the model contains pseudobonds between next neighbours (black), whose lengths are chosen in such a way that they enforce the established bond angles. Furthermore, grey sticks in fig. 2.4 represent NH–R and CO–R pseudobonds which keep the side chain bond angles and L-isomerisation of the amino acids in place. Finally, a C_α–C_α pseudobond (yellow) enforces the known distance of 3.8 Å between these two atoms and thereby the *trans* configuration of the peptide bond. The *cis* variant is neglected in the model.

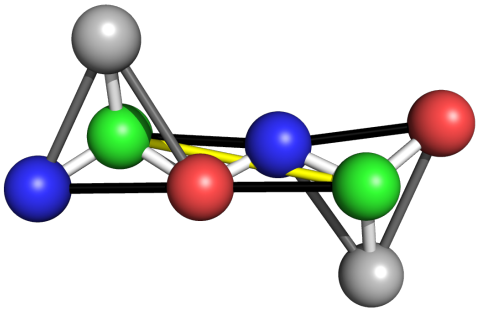

Figure 2.4 – PRIME geometry. Sizes are not to scale, bead positions are. The beads correspond to NH (blue), C_α (green), CO (red) and R (grey). Covalent bonds are represented by white sticks; pseudobonds in grey, black and yellow.

PRIME interactions

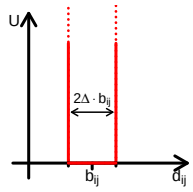

Figure 2.5 – Bond potential.

All potentials in PRIME are discontinuous. This allows the inventors to simulate the model using an efficient method called Discrete Molecular Dynamics (DMD). In Monte Carlo simulation, the discontinuous treatment is not essential but it still simplifies the calculations and therefore increases efficiency at the expense of nuance. Three types of pair potentials occur: bonds or pseudobonds, interactions between nonbonded beads, and hydrogen bonds between backbone NH and CO beads.

For the first case, the distance between bonded beads is allowed to fluctuate freely in a narrow range around an ideal bond length. As a function, this means

$$U_{\text{bond}}(d_{ij}) = \begin{cases} 0 & (1-\Delta)b_{ij} \le d_{ij} \le (1+\Delta)b_{ij} \\ \infty & \text{else} \end{cases}, \tag{2.1}$$

where d_{ij} is the distance between the (bonded) beads i and j, b_{ij} is their ideal bond length and Δ is a fluctuation parameter. This parameter was originally set to 2% [205] and later increased to 2.375% [144], which was therefore used throughout this work. The value for b_{ij} depends on the bead types (see tables 2.1 and 2.5 in section 2.4). The potential is depicted schematically in figure 2.5.

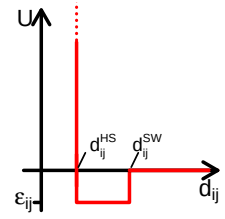

Figure 2.6 – Attractive nonbonded potential.

Two nonbonded beads experience a hard-sphere repulsion and, if both beads are side chains, an additional square well attraction:

$$U_{\text{nb}}(d_{ij}) = \begin{cases} \infty & d_{ij} < d_{ij}^{\text{HS}} \\ \epsilon_{ij} & d_{ij}^{\text{HS}} \le d_{ij} \le d_{ij}^{\text{SW}} \\ 0 & d_{ij} > d_{ij}^{\text{SW}} \end{cases}. \tag{2.2}$$

Here, d_{ij}^{HS} and d_{ij}^{SW} are the hard-sphere and square-well diameters and ϵ_{ij} the interaction strength, i.e. well depth. All of these values depend on the bead types of i and j and they are collected in section 2.4.

The interaction strength ϵ_{ij} is negative in most cases, corresponding to an attractive interaction, but positive for certain pairs of side chains, for example if both carry an equal charge.

As a special case of nonbonded interaction, the hard-sphere diameter d_{ij}^{HS} of beads separated by exactly three covalent bonds is reduced to 75% of its original value. This reduction (also called "squeeze factor") is necessary to resolve unrealistic steric clashes caused by the unification of multiple atoms into one. For example, the CO bead is composed of two atoms (C and O) of similar size. Approximating these as one bead is sufficiently accurate for nonlocal interactions but arguably not so in the short-range case. Reducing the bead size is a simple way of coping with this inaccuracy.

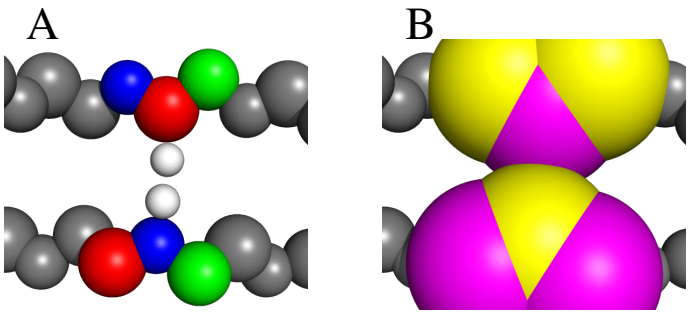

Figure 2.7 – A: Hydrogen bond geometry. During an attempt to form an H-Bond between beads NH_i and CO_j (central blue and red spheres), the H and O positions (white) are calculated from the positions of CO_{i-1}, $\mathrm{C}_{\alpha i}$, $\mathrm{C}_{\alpha j}$ and NH_{j+1}. B: Auxiliary interactions. When the H-Bond is formed, the hard-sphere diameters of NH_i–$\mathrm{C}_{\alpha j}$ and NH_i–NH_{j+1} (yellow) are increased as well as the diameters of CO_j–CO_{i-1} and CO_j–$\mathrm{C}_{\alpha i}$ (pink). These beads are depicted in their actual size, all other beads reduced to 1/4 of their diameter for improved visibility.

Finally, a hydrogen bond between the backbone amino and carboxy groups, the most important interaction for protein structure formation, is modelled as a square well potential between NH and CO beads (as in fig. 2.6) with some additional restrictions:

- Neither partner may already be involved in another H-Bond.
- The H and O positions, normally not expressed in the model, are calculated. For each of these, a line is constructed which contains the corresponding N or C position and is perpendicular to the pseudobond between the neighbouring beads. This line is extended by 1 Å for the H atom and by 1.2 Å for the O atom. Using these H and O positions, the angles ∠NHO and ∠COH can be calculated and both must be larger than 120° for the H-Bond to be formed.
- The interaction partners must be separated by at least 3 intervening residues.

The first restriction enforces the specificity of hydrogen bonds because the H atom can only form one such bond. H-Bonds in nature tend to have a straight N-H-O axis, which is reflected by the second restriction. In addition to this angular restriction, the bond is stabilized by inflating the neighbouring beads in their interactions, as seen in fig. 2.7 B: the effective hard-sphere diameters d_{ij}^{HS} between the NH partner (bottom yellow sphere) and both neighbours of the CO partner (top yellow spheres) are increased as well as d_{ij}^{HS} between CO and the neighbours of NH (pink spheres). This increased bead size, also called "auxiliary interactions", reduces the risk of breakage of the bond by bending during a DMD simulation. Such a process does not exist in MC, hence the size change is implemented as a fourth restriction on H-Bond formation. Finally, the third restriction forbids very tight turns in the chain which would probably be broken again immediately by the auxiliary interactions. If all of these restrictions are obeyed by the configuration, an H-Bond is formed and yields an energy gain of -1. All side chain energies (table 2.3) are given in relation to this H-Bond energy.

Coarse-grained simulations usually employ an implicit solvent, meaning that the solvent molecules are factored into the energy scale by a mean-field approximation instead of considering them directly.

Hence, energies and temperatures can either be expressed in physical units or as reduced, dimensionless quantities. Throughout this thesis, the symbols U, E and T will be used to refer to the reduced potential energy, total energy and temperature. The physical energies and temperatures will be named $U' = \epsilon_{\mathrm{HB}} U$, $E' = \epsilon_{\mathrm{HB}} E$ and $T' = \epsilon_{\mathrm{HB}} T / k_B$. The conversion depends solely on the effective H-Bond energy ϵ_{HB}, whose value depends on the amino acids involved in a protein. A detailed discussion of these values will take place in section 6.1; in the earlier chapters only the reduced quantities will be used.

PRIME20

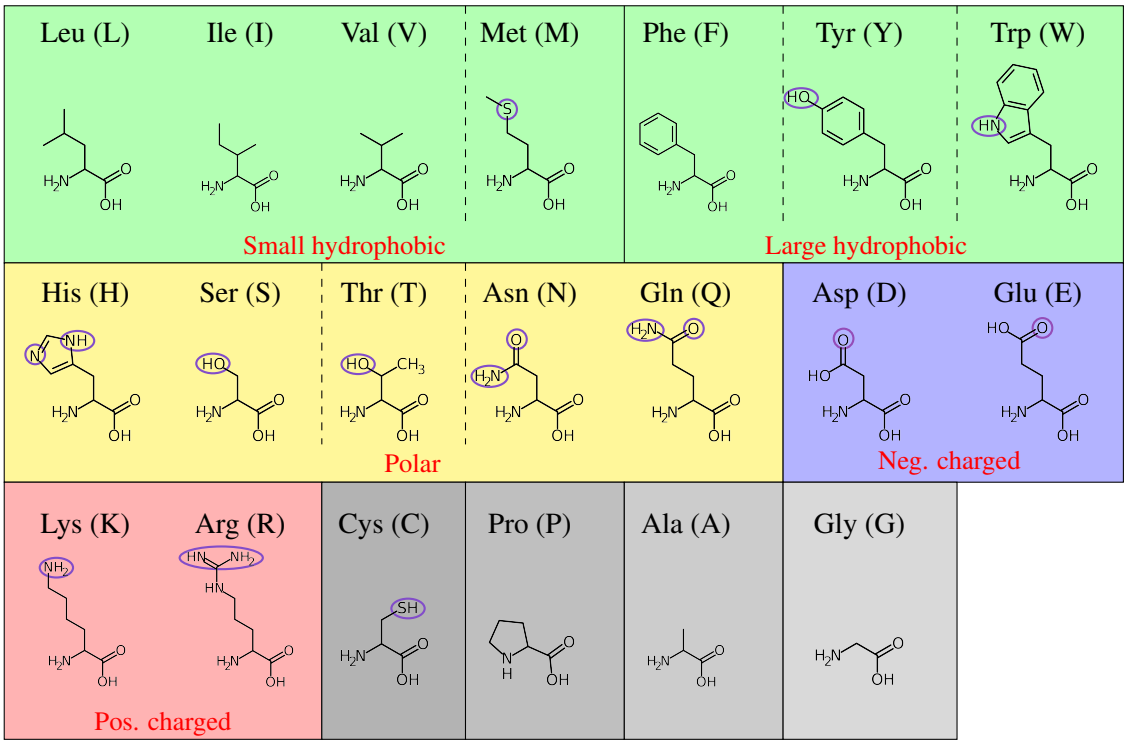

Figure 2.8 – Classification of amino acids in PRIME20. Atoms and groups capable of forming side chain H-Bonds are marked by purple ellipses.

The PRIME20 model uses the geometry and potentials of PRIME and adds individual parameters for all 20 types of side chains. Its development has been described in detail by Cheon et al. in [36]. In short, the research group selected 711 different PDB structures and calculated distance distribution histograms to model individual bead sizes and square well diameters. Square well depths were obtained by feeding these 711 structures and about two million decoys to an optimisation algorithm which modified the parameter set until most of the correct structures lay in local energy minima.

A similar optimisation has been performed by Miyazawa and Jernigan before [133, 134] and resulted in a set of $\frac{20 \cdot 21}{2} = 205$ parameters, one for each possible pair of side chains. Cheon et al. used a classification scheme depicted in fig. 2.8 to reduce the number of parameters and thereby the dimension of the minimising algorithm. As the figure portrays, most side chains can be grouped as hydrophobic, polar or carrying a charge. Special cases are cysteine due to its capability to form disulfide bonds, proline due to its overall unusual shape, alanine and glycine due to their very small side chains. In fact, glycine is modelled without a side chain in PRIME20. The hydrophobic residues are distinguished between large (defined by the aromatic ring capable of $\pi - \pi$ stacking) and small, and finally the existence of possible proton donors and acceptors for hydrogen bonds subdivides some of the classes further (marked by dashed lines). This leaves 13 groups (excluding glycine, which does not interact), reducing the number of

parameters to $\frac{13 \cdot 14}{2} = 91$. However, this is not the final number of parameters because many interactions have a very similar character and are assigned the same energy. For example, the interaction of the large hydrophobic side chains of phenylalanine, tyrosine and tryptophan with the small hydrophobic (leucine) group will be the same because the leucine, isoleucine and valine side chains do not form hydrogen bonds and therefore the polar groups of tyrosine and tryptophan are irrelevant. This means that the Phe-Leu, Tyr-Leu and Trp-Leu interactions can be described by the same parameter. Methionine on the other hand has a possible proton acceptor and therefore requires a different parameter for its interaction with tyrosine and tryptophan. Similar considerations reduced the final number of interaction parameters to just 19, which were then optimised as described above. The following section lists the resulting values used in this work.

2.4 A full set of parameters

Applying PRIME20 turned out to be somewhat challenging due to the fact that the model parameters are spread through literature and there is no single publication containing all of them at once. As of early 2019, there are 14 published papers constructing or applying PRIME [74, 122, 123, 124, 140, 141, 142, 143, 144, 164, 204, 205, 206, 207], 14 papers constructing or applying PRIME20 [36, 37, 38, 39, 40, 108, 109, 180, 211, 212, 213, 217, 218, 219], 7 PhD dissertations from C. K. Hall's group which are mostly made up of these papers but also contain additional information and discussion [107, 121, 139, 163, 210, 216, 203] and one further dissertation from outside the Hall group [174]. The model used in most of the present work is PRIME20 as described by Cheon et al. in 2010 [36] and by the sources cited there. This section provides a comprehensive list of the parameters as well as a discussion of misconceptions regarding the aforementioned squeeze factor and a description of unused parameters which can be found in literature.

Parameters of PRIME20

The first important resource is the paper by Voegler Smith and Hall from 2001 [205] in which PRIME was first described. Most of the backbone geometry is found here, including the diameters, bond and pseudobond lengths of the backbone beads listed in table 2.1 as well as the 75% squeeze factor and the H-Bond square well distance. For the interaction between two backbone beads, the arithmetic mean of their diameters is used according to the Lorentz-Berthelot combining rule [117].

Table 2.1 – PRIME backbone geometry [205]

Diameter	(Å)	Bond length	(Å)	Pseudobond length	(Å)	Other	
NH	3.3	NH_i–$C_{\alpha i}$	1.46	NH_i–CO_i	2.45	Pseudobond $C_{\alpha i}$–$C_{\alpha i+1}$	3.8 Å
C_α	3.7	$C_{\alpha i}$–CO_i	1.51	$C_{\alpha i}$–NH_{i+1}	2.41	NH$\cdots$CO square well	4.2 Å
CO	4.0	CO_i–NH_{i+1}	1.33	CO_i–$C_{\alpha i+1}$	2.45	3-bond squeeze factor*	75%

*This factor reduces the diameters of nonbonded beads which are separated by up to three bonds along the chain.

Some further development of the model is described in Nguyen, Marchut and Hall's 2004 paper [144]. Here, the allowed bond fluctuation parameter Δ (eq. (2.1)) was increased from 2% to 2.375% and the auxiliary interactions (table 2.2, cf. fig. 2.7 B) were introduced.

Table 2.2 – Auxiliary interaction distances of an $NH_i \cdots CO_j$ H-Bond [144]

NH_i–$C_{\alpha j}$	NH_i–NH_{j+1}	CO_j–$C_{\alpha i}$	CO_j–CO_{i-1}
5.00 Å	4.74 Å	4.86 Å	4.83 Å

In the first PRIME20 paper published in 2010 [36], Cheon, Chang and Hall list the side chain interaction energies and in the supporting information the side chain bead and square well diameters. They are shown here for the amino acids relevant to the present work: alanine (A), serine (S), glutamine (Q), arginine (R) and tryptophan (W). All further values can be found in appendix A.

Table 2.3 – PRIME20 side chain energies of alanine, serine, glutamine, arginine and tryptophan [36]. Positive values are coloured blue, negative values red.

	Energy (ϵ_{HB})				
	A	S	Q	R	W
A	-0.084	0.074	0.074	0.074	-0.148
S		-0.086	-0.086	-0.086	-0.086
Q			-0.080	-0.086	-0.086
R				0.073	0.015
W					-0.205

Table 2.4 – PRIME20 side chain geometry of alanine, serine, glutamine, arginine and tryptophan [36, SI].

	Bead diameter (Å)					Square well width (Å)				
	A	S	Q	R	W	A	S	Q	R	W
A	2.7	2.3	3.0	3.0	2.7	5.4	5.9	5.8	6.1	5.5
S		2.5	2.7	3.0	2.7		6.4	6.0	6.3	6.3
Q			3.6	3.6	3.4			6.6	6.9	6.7
R				3.2	3.0				7.2	6.9
W					3.7					7.4

Finally, the side chain bond and pseudobond lengths are found in the supporting information of Cheon, Chang and Hall's paper from 2015 [39]. The paper itself also contains the masses of all beads except for the cysteine, proline and tryptophan side chains. These three (as well as all other side chains, but not the backbone beads) are found in Wagoner's PhD thesis [210, Table 4.1], albeit with only two decimals where Cheon et al. use three.

Table 2.5 – Side chain bond and pseudobond lengths (in Å) and bead masses (relative to the CH_3 mass) [39]. The W side chain mass was taken from [210] instead.

	C_α–R	NH–R	CO–R	Mass		Mass
A	1.600	2.500	2.560	1.000	NH	0.999
S	1.967	2.650	2.800	2.064	C_α	0.866
Q	3.300	3.750	4.000	4.795	CO	1.863
R	4.200	4.500	4.800	6.728		
W	3.881	4.100	4.350	8.66		

Squeeze factors and PRIME20n

The above set of parameters defines the PRIME20 model as it has been used for most of the present work. Unfortunately, this model has been found at a late stage of the project to fail to produce some of the configurations which are to be expected in realistic polypeptides, and it is also not identical to the actual model version as it is being used by its creators.

The main difference between PRIME20 and the original model (which will be called PRIME20n in the following) are the unequal squeeze parameters. As will be explained in detail in chapter 5, the 3-bond squeeze factor (see table 2.1) turns out to be insufficient for the formation of α-helices, one of the essential elements of protein structure. At least one 4-bond squeeze factor is required to form this kind of structure, as also noted in Gil Rutter's PhD thesis [174], who graduated from an unrelated research group. As Yiming Wang, a recent graduate from the Hall group, confirmed in private communication, the correct model (i.e. PRIME20n) contains a total of ten squeeze factors to be applied to the interaction of bead pairs separated by 3, 4, or (in one case) 6 bonds.

These squeeze factors are divided into two groups of five. The factors in the first group (*sqz*1-5, table 2.6) apply to interactions of backbone beads. They are also found in Anne Voegler Smith's PhD thesis [203, source code on p. 256], indicating that they are used in PRIME as well and that the factor of 75% given in the papers is an approximation. The second group (*sqz*6-10, table 2.7) consists of squeeze parameters applying to interactions between side chain and backbone beads. These are also found in the SI of Cheon, Chang and Hall's 2015 paper [39]. They were originally not implemented because the table caption (which calls them "minimum distances between neighboring spheres not having covalent or pseudo-bonds") was misunderstood to describe a simulation result rather than a model parameter.

Table 2.6 – Backbone squeeze factors *sqz*1-5 [203], original and squeezed bead diameters.

	*sqz*1 $C_{\alpha i}$–CO_{i+1}	*sqz*2 $C_{\alpha i}$–NH_{i-1}	*sqz*3 CO_i–NH_{i+2}	*sqz*4 NH_i–NH_{i+1}	*sqz*5 CO_i–CO_{i+1}
Squeeze factor	1.1436	0.88	0.87829	0.8	0.7713
Original diameter (Å)	3.85	3.5	3.65	3.3	4.0
Squeezed diameter (Å)	4.40286	3.08	3.074015	2.64	3.0852

Table 2.7 – Side chain squeeze diameters *sqz*6-10 [39] for the relevant amino acids. Values in Å.

	*sqz*6 CO_{i-1}–R_i	*sqz*7 NH_{i+1}–R_i	*sqz*8 $C_{\alpha i+1}$–R_i	*sqz*9 $C_{\alpha i-1}$–R_i	*sqz*10 CO_{i-2}–R_i
A	4.598	3.312	3.000	4.353	4.997
S	4.507	3.331	3.128	4.380	4.944
Q	5.134	4.139	3.996	5.062	5.000
R	5.703	4.827	4.651	5.535	4.978
W	5.180	4.460	4.187	4.963	4.986

In all other cases, the effective hard-sphere repulsion distance between side chain and backbone beads is assumed to be the arithmetic mean of the backbone bead diameter and the side chain bead diameter according to the Lorentz-Berthelot combining rule [117]. This method requires the existence of a single bead diameter for a given side chain, a parameter which is not found in table 2.4 as the values there have been determined for pairs of beads. Any attempt to deconvolute the table by applying the Lorentz-Berthelot rule backwards fails. The best choice for a side chain bead diameter appears to be the self-interaction value, found on the diagonal of table 2.4.

An alternative method of assigning side chain bead diameters could be based on a comparison of the squeezed diameters in table 2.7 to the respective squeeze factors. Assuming that (1) every side chain has a well-defined diameter to be used for interactions with the backbone, and (2) the squeezed diameters are derived from this value by applying constant squeeze factors like in the backbone case (*sqz*1-5), the side chain diameter can easily be extracted if these squeeze factors are known. The procedure can be performed independently for all five squeezed interactions and should – if these assumptions hold – yield the same result five times.

The requirement of constant squeeze factors appears to be fulfilled as they are found alongside *sqz*1-5 in the appendix of Voegler Smith's PhD thesis [203]. The five factors and the resulting diameters are shown in table 2.8. Evidently, the diameters for any single side chain are far from equal, hence one of the assumptions must be wrong. Either such a well-defined diameter does not exist or these squeeze factors are only valid in PRIME20, not in PRIME20n. In either case, the self-interaction diameter remains the best guess. It is however interesting to note that all values in table 2.8 are considerably larger than the self-interaction diameters.

Table 2.8 – Side chain squeeze factors *sqz*6-10 [203] and "original" bead diameters of alanine, serine, glutamine, arginine and trypthophan side chains obtained by inverse application of the combining rule. Squeeze factors are dimensionless, diameters in Å and rounded to three decimals.

	*sqz*6 CO_{i-1}–R_i	*sqz*7 NH_{i+1}–R_i	*sqz*8 $C_{\alpha i+1}$–R_i	*sqz*9 $C_{\alpha i-1}$–R_i	*sqz*10 CO_{i-2}–R_i
sqz	0.7607	0.7930	1.0956	1.1244	0.9259
A	4.708	4.266	4.246	4.479	6.794
S	4.758	4.589	4.296	4.317	6.679
Q	6.882	6.778	5.541	5.432	6.800
R	8.691	8.430	6.404	6.444	6.753
W	7.726	7.260	5.360	5.514	6.770

Another difference between PRIME20 and PRIME20n is the increase of the backbone H-Bond distance (the NH$\cdots$CO square well) from 4.2 Å to 4.5 Å, which was mentioned in later publications [38, 211, 212, 213].

Further parameters

Several details of PRIME20(n) are documented in the publications of the Hall group, but have not been used in this work. They are listed here for the sake of completeness and because some of them are relevant for future use of the model.

- The correct behaviour of the auxiliary interactions (table 2.2) is somewhat unclear. While they appeared to represent increased hard-core repulsion distances in the first papers [205], Nguyen et al. describe their effect as a temporary square shoulder repulsion event [144]. In the present work, the auxiliary distances were implemented as an additional constraint in H-bond formation, which is more alike to Voegler Smith's description. The process described by Nguyen et al. would cause ambiguity in energy calculation.

- An eleventh squeeze parameter reduces the size of two NH and CO beads which are interacting in an H-Bond from originally 3.65 Å to 2.352 Å (Y. Wang, private communication). The reduction only applies to the interaction of the H-bond partners. Implementing this parameter in a MC simulation is virtually impossible because, like the auxiliary interactions, it creates ambiguous situations not only regarding the energy, but even regarding the legality of a given conformation.

- The proline backbone amide group cannot form H-Bonds. This fact is reflected in the model, as noted by Wang and Hall in 2016 [219]. Its implementation has not been necessary yet as no proline-containing peptides have been simulated, but it will be used in the future.

- The model contains an energy parameter for the H-Bond interaction between eligible side chain and backbone beads. The interaction energy is very small (-0.015), even in comparison to the weak side chain attractions used in the present work. In a later publication [39, SI], its value was

even reduced to 0. Coupled with the fact that square well diameters for this interaction are not documented and that it requires considerable additional computation time, it was ignored in the present work, but some results suggest that it may be useful for future simulations.

- Cheon et al. proposed a double well potential in 2015 to replace the square well interaction of side chains [39], having felt the need for a more detailed representation. A comparison of chain behaviours applying the different potentials might be quite instructive in the future, but has not been performed yet.

- Marchut et al. simulated polyglutamine using a modification of PRIME with four side chain beads rather than one [122, 123, 124]. In this modification, they also increased the bond fluctuation parameter to 2.5% and left out the restriction for H-Bonds to be formed only between beads with at least three intervening amino acids. The group revoked these changes during the extension to PRIME20, thus making them irrelevant for current application.

- In another "sidetrack", Cheon et al. introduced a modified set of auxiliary interactions [38] by which parallel β-sheets are preferred over the antiparallel geometry. These specific interactions are not of interest here.

- In 2016, Cheon et al. also published another set of side chain geometry parameters to make the peptide more flexible. Although apparently successful in MD simulation, this modification appears to have been used only temporarily.

- Finally, the above tables (2.3-2.8) only show values for five side chains. Values for all twenty side chains can be found in appendix A (tables A.1-A.6).

2.5 FRET and TTET

In chapter 6, a comparison between simulations and experimental results obtained by Peter Enke and Michael Schleeger will be discussed. They used Förster resonance energy transfer (FRET) and Triplet-triplet energy transfer (TTET) to observe the conformational ensembles of polyglutamine, polyserine and other peptides. Because the experiments require some modifications to the primary structures, it is useful to adapt these to the simulation with PRIME20 as well.

In both FRET and TTET [62], two chromophores are added at certain positions along the molecule – in the cases considered here always at the chain termini. Although the underlying mechanisms of energy transfer differ clearly, the methods are very similar from a coarse-grained simulation point of view: a laser pulse excites one of the chromophores, the energy is partially transferred to the second chromophore in a process whose probability and reaction rate depend on the distance, and finally the second chromophore returns to its electronic ground state, emitting a photon to be detected. Typical results of these experiments are data regarding the chain dynamics, which a Monte Carlo simulation cannot reproduce, but also average contact probabilities and distance histograms between the chromophores in thermodynamic equilibrium. These quantities can and will be compared in chapter 6.

If the comparison between simulation and experiment is to be meaningful, the experimental set-up should be reproduced as accurately as possible within the limitations of the model. This means that the chromophores depicted in fig. 2.9 need to be represented somehow. PRIME20 provides parameters for all 20 proteinogenic amino acids, but these chromophores are usually different synthetic molecules and therefore not included in PRIME20. Adding them to the model as new side chains is not possible in a mathematically rigorous way because the parameter optimisation is based on the knowledge of folded structures which would not be available for these unusual residues. For this reason, the chromophores

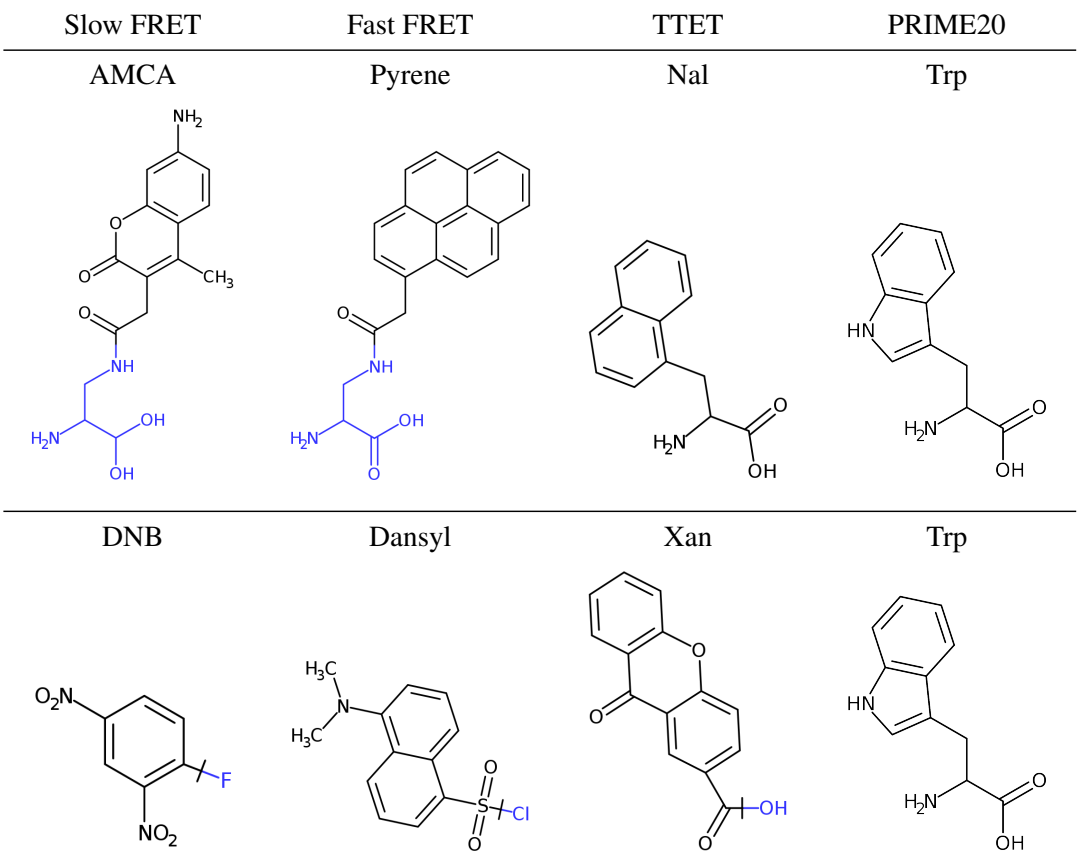

Figure 2.9 – FRET and TTET chromophores and tryptophan, used as a "chromophore" in PRIME20. Two pairs of FRET chromophores with different lifetimes (hence "slow" and "fast") and one TTET pair are used. AMCA and Pyrene are attached to the chain via a peptide linker; Naphthylalanine (Nal) is a nonproteinogenic amino acid. DNB, Dansyl and Carboxyxanthone (Xan) are attached to the N-terminus at their respective marked reaction sites.

must be approximated by proteinogenic amino acids. Since all of the chromophores (fig. 2.9) are relatively large in comparison to the proteinogenic amino acids and contain at least one aromatic ring, the obvious choice is to use the largest amino acid available, tryptophan, which has in fact even been used as a FRET donor elsewhere [214].

Chapter 3

SAMC

As stated at the beginning of the previous chapter, the algorithm used in this work is called SAMC, short for Stochastic Approximation Monte Carlo. The present chapter introduces the method, briefly leading from the general Monte Carlo concept via the famous Metropolis Monte Carlo to SAMC and related methods. The later sections describe technical details of the implementation and introduce physical observables to be used in the results chapters.

3.1 Standard Monte Carlo methods

The aim of any simulation, broadly speaking, is to measure observables which depend on the configuration of a system and characterise aspects of its behaviour. Let A be any observable, disregarding its physical significance for now. (Relevant observables are listed in section 3.4.) The average of A in a generic statistical ensemble would look like this:

$$\langle A \rangle = \sum_{\mathbf{x}} A(\mathbf{x}) p(\mathbf{x}) \,, \tag{3.1}$$

where $\mathbf{x}$ is an element of configuration space (i.e. a set of coordinates) and $p(\mathbf{x})$ a weight function, for example – in the canonical ensemble – the Boltzmann weight at a given temperature T:

$$p(\mathbf{x}, T) = \frac{e^{-H(\mathbf{x})/k_B T}}{Z(T)} \,. \tag{3.2}$$

Here, $H(\mathbf{x})$ is the Hamilton function and its value is the total (here: configurational) energy of the system. T is the temperature, k_B the Boltzmann constant. The partition function is defined as $Z(T) = \sum_{\mathbf{x}} e^{-H(\mathbf{x})/k_B T}$. It should be noted that the sums in all following equations as well as this and eq. (3.1) are often written as integrals instead, but for most simulations the discrete notation using sums is more practical.

If all possible $\mathbf{x}$ can be listed systematically (and if their number is small enough), eq. (3.1) is easily evaluated. However, this is not the case in an off-lattice model, so only a subset of configuration space can be sampled. According to the law of large numbers, the average of A over this subset tends towards the ensemble average as the sample size tends towards infinity:

$$\langle A \rangle' = \sum_{t=1}^{n} A(\mathbf{x}_t) p(\mathbf{x}_t) \xrightarrow[n \to \infty]{} \langle A \rangle \,. \tag{3.3}$$

The selection of such a sample from configuration space usually relies on either of two basic principles: Molecular Dynamics (MD) or Monte Carlo (MC). An MD simulation attempts to emulate the physical behaviour of particles by calculating forces between them from their potential functions, velocities from the forces and displacing the particles according to these velocities. MD ideally produces

a realistic causality of force and movement, but obtaining a reasonable sample of configuration space requires considerable computational power, which, even after decades of exponential growth [136], is still a limiting factor.

This issue is addressed by Monte Carlo techniques: in MC, forces are ignored and the particles are displaced randomly. MC can sample configuration space much faster than MD, but since velocities and forces are ignored, dynamic quantities are usually impossible to access, hence the amount of available information is reduced. MC simulations are often performed using coarse-grained models while atomistic models are more common in MD to make best use of the respective advantages.

The basic MC procedure, called simple sampling, involves creating a sequence $(\mathbf{x}_t)$ of uniformly distributed random configurations, calculating $A(\mathbf{x}_t)$ and $p(\mathbf{x}_t)$ for each of them and averaging these to an approximation of $\langle A \rangle$ following eq. (3.3). The method may be viable for very simple systems, but already for the shortest peptide chains in PRIME20 it must fail because a configuration is defined by at least 72 coordinates (assuming a 6-residue peptide with 4 beads per residue and 3 cardinal coordinates for each bead) and an exponential function (see eq. (3.2)) with such a high dimension is practically equal to zero for most $\mathbf{x}$, so the majority of the simulation will consume time without really contributing to the calculation of $\langle A \rangle$.

To increase efficiency, configuration space must be visited non-uniformly with a probability distribution equal or similar to $p(\mathbf{x})$. In this case, called importance sampling, $\langle A \rangle$ is approximated by $\sum_t A(\mathbf{x}_t)$ and because points of high importance are selected more frequently than those with low $p(\mathbf{x})$, the simulation converges faster. Unfortunately, importance sampling in the canonical ensemble is not trivially done because $Z(T)$ as a normalisation factor is not known *a priori*, hence $p(\mathbf{x})$ cannot be calculated. The solution to this problem has been published by Metropolis et al. in 1953 and is now known as Metropolis or Markov Chain MC [97, 131]. Although the method itself has not been used in the present work, it can be considered as the foundation of most current MC methods and its knowledge helps to understand advanced methods like SAMC.

The "trick" of Metropolis MC lies in the observation that the unknown $Z(T)$ cancels out of the quotient

$$\frac{p(\mathbf{x}_1, T)}{p(\mathbf{x}_2, T)} = \frac{e^{-H(\mathbf{x}_1)/k_B T}}{Z(T)} \frac{Z(T)}{e^{-H(\mathbf{x}_2)/k_B T}} = \frac{e^{-H(\mathbf{x}_1)/k_B T}}{e^{-H(\mathbf{x}_2)/k_B T}} \,. \tag{3.4}$$

With this equation in mind, a process is generated in which the selection of a configuration depends on the previous configuration (called a Markov process). Mathematically, this process is derived from the Master equation (3.5). It sums up all probability flows between any pair of states $(\mathbf{x}_i,\ \mathbf{x}_j)$ during one simulation step $(t-1 \to t)$ and postulates a conservation of probability:

$$p_t(\mathbf{x}_i) = p_{t-1}(\mathbf{x}_i) + \sum_{j \neq i} p_{t-1}(\mathbf{x}_j) w(\mathbf{x}_i|\mathbf{x}_j) - \sum_{j \neq i} p_{t-1}(\mathbf{x}_i) w(\mathbf{x}_j|\mathbf{x}_i) \,. \tag{3.5}$$

$w(\mathbf{x}_i|\mathbf{x}_j)$ signifies the conditional probability to reach state $\mathbf{x}_i$ in step t if the previous state was $\mathbf{x}_j$, synonymous to the transition probability from $\mathbf{x}_j$ to $\mathbf{x}_i$. The T-dependence of p can be left out of the equation because temperature is a constant parameter in a Metropolis simulation. Furthermore, $p(\mathbf{x})$ is independent of t in thermodynamic equilibrium, simplifying the equation to

$$\sum_{j \neq i} p(\mathbf{x}_j) w(\mathbf{x}_i|\mathbf{x}_j) = \sum_{j \neq i} p(\mathbf{x}_i) w(\mathbf{x}_j|\mathbf{x}_i) \,. \tag{3.6}$$

The algorithm to be designed must fulfil this condition. Because the sums are uncomfortable to handle, they are removed by imposing detailed balance:

$$\frac{p(\mathbf{x}_i)}{p(\mathbf{x}_j)} = \frac{w(\mathbf{x}_i|\mathbf{x}_j)}{w(\mathbf{x}_j|\mathbf{x}_i)} \quad \forall\, j \neq i \,. \tag{3.7}$$

If the algorithm adheres to this condition, eq. (3.6) is satisfied as well. However, it is also possible to construct advanced MC algorithms which ignore detailed balance and only follow the Master equation [24, 91].

Any algorithm whose transition probabilities $w(\mathbf{x}_i|\mathbf{x}_j)$ comply with this eq. (3.6) will produce the desired probability distribution $p(\mathbf{x_i})$. The most commonly used option is

$$w(\mathbf{x}_i|\mathbf{x}_j) = \min\left(1, \frac{p(\mathbf{x}_i)}{p(\mathbf{x}_j)}\right) . \tag{3.8}$$

If $p(\mathbf{x})$ are the Boltzmann weights (eq. (3.2)), this corresponds to the fraction of probabilities in eq. (3.4) and therefore can be calculated in ignorance of $Z(T)$.

For the implementation of Markov Chain MC, each step is further divided into a proposal and an acceptance portion, $w(\mathbf{x}_i|\mathbf{x}_j) = \text{prop}(\mathbf{x}_i|\mathbf{x}_j) \cdot \text{acc}(\mathbf{x}_i|\mathbf{x}_j)$. During proposal, a configuration $\mathbf{x}_i$ is created, usually by modifying a part of the current configuration $\mathbf{x}_j$. Many different algorithms for this modification (Monte Carlo moves) have been published, some of which will be discussed in section 3.3. The main condition for all of these moves is microscopic reversibility, i.e. the postulation that $\text{prop}(\mathbf{x}_i|\mathbf{x}_j) = \text{prop}(\mathbf{x}_j|\mathbf{x}_i)$ for all (i, j). This way, the proposal probabilities cancel out of eq. (3.7) and the acceptance portion remains, during which either the newly proposed $\mathbf{x}_i$ or the old $\mathbf{x}_j$ is selected as the next configuration with probabilities according to eq. (3.8).

3.2 Flat-Histogram Monte Carlo and SAMC

Since the publication of the famous paper by Metropolis et al. 66 years ago [131], several more advanced MC techniques have been developed, among them Stochastic Approximation Monte Carlo (SAMC). The motivation for SAMC and related methods lies in two observations about the old Metropolis algorithm. Firstly, as mentioned, $Z(T)$ cancels out of the equations. The Metropolis method relies on this fact, but knowledge of the partition function would be valuable as the entire canonical thermodynamics can be derived from it: the free energy $F(T) = -k_B T \ln[Z(T)]$ is a fundamental quantity and with knowledge of $Z(T)$, eq. (3.1) is easily evaluated as well. Secondly, if $p(\mathbf{x}, T)$ has multiple distinct maxima (which is a defining characteristic of first-order phase transitions), a simulation at low temperature can get trapped in one of the respective states and will be unable to produce a correct thermodynamic average. Both issues are addressed by the class of Flat Histogram Monte Carlo methods.

The idea of Flat Histogram MC is to circumvent the issue of multiple maxima in $p(\mathbf{x}, T)$ by essentially ignoring the maxima and producing an uniform, flat histogram of visited states. Since the probability weights are unknown *a priori*, this is achieved by an iterative approximation scheme, in which the acceptance probability depends in some way on the current estimate of $p(\mathbf{x}, T)$. A successful simulation will therefore yield these weights and consequently also $Z(T)$. Depending on algorithmic details, these methods are called Umbrella sampling [196], Multicanonical MC [75, 88, 146], Well-tempered ensemble metadynamics [30], Wang-Landau [28, 29, 88, 215, 223, 224] or SAMC [88, 113, 114, 115, 179, 222].

The aim of an SAMC simulation is to approximate the microcanonical configurational density of states $g(U)$. If $g(U)$ is known, the energy-dependent Boltzmann weights can be calculated as

$$p(U, T) = g(U) e^{-U/k_B T} , \tag{3.9}$$

producing the partition function

$$Z(T) = \sum_U g(U) e^{-U/k_B T} \tag{3.10}$$

and the canonical ensemble average of an observable A

$$\langle A \rangle(T) = \frac{1}{Z(T)} \sum_U A(U) g(U) e^{-U/k_B T} . \tag{3.11}$$

The algorithm is depicted in a slightly simplified form as a flowchart in fig. 3.1. The simulation begins with an estimate of $g(U)$, often $\tilde{g}(U) \equiv 1$, a visit histogram $H_{\text{vis}}(U) \equiv 0$ and a modification factor γ_0. Given a configuration $\mathbf{x}_1$ with energy U_1, a new configuration $\mathbf{x}_2$ with energy U_2 is proposed by means of any Monte Carlo move like in Markov Chain MC. The acceptance probability is the inverse of the current estimate of $g(U)$:

$$\text{acc}(U_2|U_1) = \min\left(1, \frac{\tilde{g}(U_1)}{\tilde{g}(U_2)}\right) . \tag{3.12}$$

Thus, states with high $\tilde{g}(U)$ will be visited less frequently than states with low $\tilde{g}(U)$. Because the proposed $\mathbf{x}$ are uniformly distributed, their distribution depending on U is equal to $g(U)$. Combined with the acceptance probability – assuming $\tilde{g}(U)$ is a good approximation of $g(U)$ – the total number of visits will be equal for all U. A flat visit histogram indicates uniform visit probabilities and thereby a good quality of $\tilde{g}(U)$, hence the term flat histogram MC.

To reach this flat histogram, the initial $\tilde{g}(U)$ must be modified during simulation. In SAMC, $\tilde{g}(U)$ is updated after each MC step according to

$$\ln[\tilde{g}_{t+1}(U)] = \ln[\tilde{g}_t(U)] + \gamma_t \delta_{U,U_{\text{new}}} + \boldsymbol{\pi}(U) , \tag{3.13}$$

where $\boldsymbol{\pi}(U)$ is a bias vector which can be used to enhance low visit numbers and U_{new} signifies the energy of the state U_1 or U_2 selected in the acceptance step. γ_t usually follows the sequence

$$\gamma_t = \gamma_0 \min\left(1, \frac{t_0}{t}\right) . \tag{3.14}$$

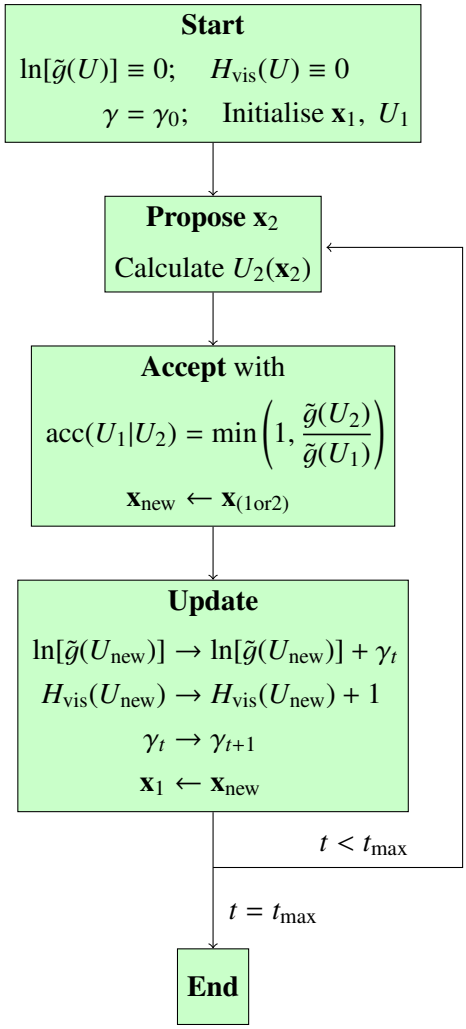

Figure 3.1 – Flowchart of the SAMC algorithm without bias $\boldsymbol{\pi}(U)$.

Other sequences are possible according to certain criteria: $\sum_t \gamma_t$ must diverge and $\sum_t \gamma_t^\zeta$ must converge for some $\zeta \in (1, 2)$ [113]. The rigorous formulation of these criteria is the main difference between SAMC and Wang-Landau MC, in which γ is only reduced after $H_{\text{vis}}(U)$ reaches a uniform distribution. This algorithm is more intuitive because it "waits" for convergence, but it has been proven to leave a nonzero residual error [178] and the time to reach a desired $\gamma_{\min}$ is unpredictable, contrary to SAMC, which does converge and reaches $\gamma_{\min}$ within a pre-set number of steps.

The parameters γ_0 and t_0 are used to improve the convergence behaviour. If $\boldsymbol{\pi}(U)$ is not constant, the resulting visit histogram will be proportional to it and the bias has to be corrected from $\tilde{g}(U)$ at the end.

3.3 Implementation details and MC moves

The SAMC simulation used in this project was written in C and is based on an earlier program by Mark P. Taylor and Benno Werlich used to simulate one-bead homopolymers. Due to the requirements of the four-bead geometry, the Monte Carlo moves (see below) differ from the ones found there [222]. The calculation of bead overlaps and energies was performed using an efficient hierarchical scheme described by Johnson et al. [89]. As seen in the previous chapter (esp. table 2.3), interaction energies are provided

with three decimals, but only certain values can actually occur. Therefore the energy scale is divided into bins of width 0.1. Statistical analysis and most graphics were produced in R [167] and all depictions of chain configurations were drawn using the Python-based software PyMOL [177].

Three types of MC moves were implemented for this model: a **local displacement**, a **pivot rotation**, and a **configurational bias** or cut-and-grow method. Among these, the **local displacement** occurs most frequently during simulation (about N times as often as the pivot move). In this move, the position of a single bead is changed slightly and randomly. The change is very small and usually does not even affect the energy of the chain. It is however necessary in order to fully sample configuration space because moves like the pivot rotation are unable to change local distances and bond lengths.

As a detail, it may be worth noting that the move was originally written in spherical coordinates (randomly and uniformly selecting a direction on the unit sphere and a distance), but this method turned out to cause a bias in the resulting $\tilde{g}(U)$, so it was changed to Cartesian coordinates (selecting three independent random distances along the coordinate axes) instead. The only parameter to be tuned is the maximum displacement distance, typically 0.01 Å. A larger distance leads to an increased impact on the configuration but also to decreased acceptance rates, mostly due to violations of bond length constraints.

The second move is the **pivot rotation**. Here, one dihedral angle is selected and its value changed randomly, which results in a rotation of one end of the molecule about this pivot bond. The move does not change bond lengths like the local displacement, instead it affects the configuration on a global scale, which is essential for an efficient MC simulation. Due to the larger impact, its acceptance rates are much lower, usually about 13%, whereas about 75% of local move attempts were accepted. On the technical side, the pivot move is handled differently depending on whether it affects the Φ or Ψ angle, and a pivot rotation about the peptide bond was not implemented. The only relevant parameter for the pivot move, the maximum rotation angle, was usually set to $\pi/3$ in both cases (Φ, Ψ). As another parameter, it would have been possible to limit or bias the eligible bonds to be rotated about, but in such a design, microscopic reversibility is not evident and needs to be verified, which never appeared to be worth the effort.

As an alternative or a supplement to the pivot move, a rotation affecting an interior chain fragment rather than one end was considered too. Rotation of one bead about the axis defined by its neighbours (as used in homopolymer simulation [222]) is not viable due to the restrictions of the four-bead model, but a seven-residue concerted rotation [52] was considered and ultimately ruled out due to its computational requirements. Another interesting move not implemented is the event chain algorithm [24, 91], in which displacement of one bead initiates a series of collisions until a certain total distance has been reached. This move could have acted as an extension to the local displacement but did not seem worth the computational and programming effort either.

Moves like the local displacement are physically realistic, i.e., they represent movements which the molecule can perform similarly in nature or in MD simulation. The MC method also allows the use of unphysical moves, where major conformation changes are generated in an unrealistic way, for example by changing the connectivity of certain beads [95, 127, 154] or by chain segments passing through each other (possible during pivot rotations). These moves usually help to sample configuration space more efficiently. Another such move, which was used in an early version of the present simulations, is the **configurational bias MC** (CBMC) or cut-and-grow move [17, 68]. Invented by Rosenbluth and Rosenbluth [173], the idea of this move is to cut off a part of the molecule and let it regrow bead by bead to reach a new configuration. This move is especially powerful due to its ability to leave states in which a chain end is trapped and incapable of any realistic movement.

In CBMC, for every bead to be regrown, k_{RG} positions are proposed and one of them is selected randomly. The number of legal positions (i.e. positions which do not overlap with another part of the chain) is limited by the configuration of surrounding beads, which introduces an implicit bias towards dense configurations (explaining the name) because a lower number of legal positions means a higher probability for each of them to be selected. The bias is resolved by tracking the number of allowed

positions during every growth step and virtually regrowing the previous configuration in the same way. A comparison between the two growth processes quantifies the bias which can then be compensated during the acceptance step. Tunable parameters in CBMC are k_{RG} and the maximum number of residues to be cut off and regrown.

In the PRIME20 simulations, this move produced a further bias which was not caught by the acceptance correction and whose source is still unidentified. Because the computational cost also turned out to be unnecessarily high in comparison to the pivot rotation, the CBMC move was not used in later simulations. This means that the results presented in the following chapters were solely produced using the pivot and local displacement moves, but for larger systems it may be more useful again (assuming the bias issue is resolved).

3.4 Relevant observables

The SAMC algorithm described in the previous sections produces the configurational density of states, $g(U)$, by approximation. Given this function, a production run with fixed $g(U)$ is performed to calculate U-dependent averages of various observables. The observables calculated and used in the following chapters are listed here.

Three ensembles

As already noted, knowledge of $g(U)$ allows the calculation of averages in the T-dependent canonical ensemble as well. However, care must be taken in the interpretation because both ensembles are usually defined in phase space rather than configuration space. A transition from the configurational (N, V, U) to the full microcanonical ensemble ($N, V, E = U + K$ with the kinetic energy K) is therefore of use in some situations. As derived by Shakirov et al. [179], the full microcanonical density of states $g(E)$ can be calculated analytically if $g(U)$ is known:

$$g(E) \propto \sum_U (E - U)^{f/2-1} g(U) \Theta(E - U) \, . \tag{3.15}$$

Here, Θ designates the Heaviside step function and f the number of degrees of freedom of the system. In a PRIME20 N-mer, it is $f = 12 \cdot N - 3$ because each of the 4 beads per residue can move in 3 cardinal directions and three coordinates are restricted by conservation of momentum. As was the case for $g(U)$, a normalising factor of $g(E)$ remains unknown, which is why the above formula is written as a proportionality relation instead of an equation.

Given $g(E)$, any observable $A(U)$ can be transformed according to

$$A(E) = A(U) p(U, E) \tag{3.16}$$

with

$$p(U, E) = \frac{(E - U)^{f/2-1} g(U) \Theta(E - U)}{\sum_U (E - U)^{f/2-1} g(U) \Theta(E - U)} \, . \tag{3.17}$$

The extension of the configurational to the full microcanonical ensemble allows two treatments of the canonical ensemble too. In the configurational case, the average of an observable A is calculated by the sequence of equations:

$$Z(T) = \sum_U g(U) e^{-U/k_B T} \tag{3.18}$$

$$p(U, T) = \frac{1}{Z(T)} g(U) e^{-U/k_B T} \tag{3.19}$$

$$\langle A \rangle (T) = \sum_U A(U) p(U, T) \, . \tag{3.20}$$

For evaluation in the full canonical ensemble, U needs to be replaced with E, but the equations are unchanged.

Thus, four different statistical ensembles can be considered independently: the configurational and full microcanonical ensembles and the respective canonical ones. The number is effectively reduced to three again because the two canonical ensembles do not differ significantly, as illustrated by fig. 3.2. Subfigures A and B show a comparison of temperatures (3.2 A) and squared radii of gyration (3.2 B), as two common observables, between the configurational (black) and full (red) microcanonical ensembles. Ignoring any physical significance for now, it is evident that the graphs look wildly different and, albeit derived from the same raw data, each of them carries its own kind of information. Both microcanonical ensembles will be used in the results chapters.

In the canonical ensembles (subfigures C and D), the graphs look more alike. The canonical heat capacity (3.2 C) is shifted upwards by $f/2$ in the full canonical ensemble compared to the configurational canonical ensemble, corresponding to an expectable kinetic energy gain of $k_B T/2$ per degree of freedom, but the shape of both curves is essentially identical. The squared radius of gyration (3.2 D) is altogether unaffected by the inclusion of kinetic energy. The picture is similar for all other observables, hence a distinction between the full and configurational canonical ensembles is not useful and they will be treated as one in the results chapters.

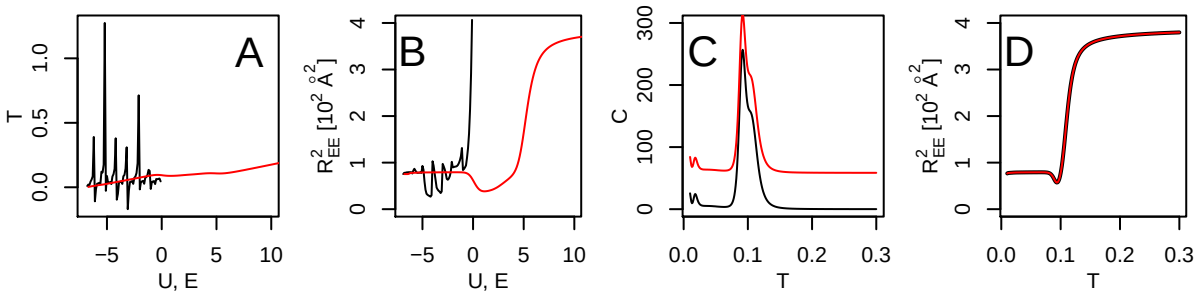

Figure 3.2 – Comparison of statistical ensembles for a few observables. A: Temperature in the configurational (black line) and full microcanonical ensemble (red line); B: Squared radius of gyration in both ensembles; C: Heat capacity in the configurational (black) and full (red) canonical ensemble, D: Squared radius of gyration in both ensembles.

Thermodynamic observables

With a definition of the three (configurational, full microcanonical and canonical) ensembles and the transformations between them (eqs. (3.17) and (3.19)), the quantities to be used in the results chapters are listed here for reference.

In microcanonical thermodynamics, the relevant observables are the density of states $g(E)$ or the Boltzmann entropy

$$S(E) = k_B \ln[g(E)] \tag{3.21}$$

and its first two derivatives, the temperature and heat capacity

$$T(E) = \left(\frac{\partial S}{\partial E}\right)^{-1} \tag{3.22}$$

$$c(E) = \left(\frac{\partial T}{\partial E}\right)^{-1} = -\frac{1}{T^2}\left(\frac{\partial^2 S}{\partial E^2}\right)^{-1} . \tag{3.23}$$

If E is replaced by U, the respective configurational quantities are calculated instead, but the equations are otherwise identical.

In canonical thermodynamics, the most relevant functions are the internal energy $U_c(T) = \langle E \rangle(T)$ and the heat capacity

$$C(T) = \frac{\langle E^2 \rangle - \langle E \rangle^2}{k_B T^2} = \frac{\partial U_c}{\partial T} \,. \tag{3.24}$$

Further quantities like the partition function $Z(T)$, the free energy $F(T) = -k_B T \ln[Z(T)]$ or the Boltzmann probabilities $p(E, T)$ are often seen in polymer physics but do not occur here.

Geometric observables

A ubiquitous geometric quantity is the squared radius of gyration

$$R_g^2 = \frac{1}{M} \sum_i m_i (\mathbf{r}_i - \mathbf{r}_{\mathrm{COM}})^2 \,, \tag{3.25}$$

where $\mathbf{r}_i$ and m_i are the position and mass of the i-th bead, $M = \sum_i m_i$ the total mass of the molecule and $\mathbf{r}_{\mathrm{COM}} = \frac{1}{M} \sum_i m_i \mathbf{r}_i$ the center of mass. In this case the index i identifies a single bead although in all other situations an index refers to a full amino acid.

By decomposing $\mathbf{r}_i$ into its components x_i, y_i and z_i, eq. (3.25) can be applied to construct the gyration tensor as a 3x3 matrix. Using the eigenvalues λ_{1-3} of this matrix the shape anisotropy κ^2 [194, 208] is defined as

$$\kappa^2 = 1 - 3 \frac{\lambda_1 \lambda_2 + \lambda_1 \lambda_3 + \lambda_2 \lambda_3}{\left(R_g^2\right)^2} \,. \tag{3.26}$$

The values of κ^2 range from 0 to 1, 0 corresponding to a spherical configuration ($\lambda_1 = \lambda_2 = \lambda_3$), 0.25 to a flat disc ($\lambda_1 = \lambda_2;\ \lambda_3 = 0$) and 1 to an idealised rod ($\lambda_2 = \lambda_3 = 0$).

Another notable geometric observable is the end-to-end distance

$$R_{\mathrm{EE}} = |\mathbf{r}_{\mathrm{NH}_1} - \mathbf{r}_{\mathrm{CO}_N}| \tag{3.27}$$

between the NH bead of the first amino acid and the CO bead of the last amino acid in the chain.

FRET

The comparison to spectroscopy experiments (FRET and TTET) has already been mentioned in section 2.5. This comparison will mostly be based on the histogram of chromophore distances

$$d = |\mathbf{r}_{\mathrm{R}_j} - \mathbf{r}_{\mathrm{R}_k}| \,, \tag{3.28}$$

where R_j designates the j-th side chain bead and j and k are the positions of the two chromophores. Importantly, this is not equal to R_{EE} even if $j = 1$ and $k = N$ because R_{EE} is the distance between NH and CO beads while d relates to side chain beads.

The FRET experiment yields a distribution function $p(d)$ which is the probability for the two chromophores to be found in an interval $[d, d + \delta d]$. This function is named $p(d, T)$ or $p_T(d)$ if evaluated at a fixed temperature. In the random coil state, the function can be fitted using the Edwards model for generic polymers with excluded volume [59, 62]. This model produces a skewed Gaussian distribution of the form

$$p_T(d) = 4\pi d^2 e^{-\left(\frac{d-b}{\sigma}\right)^2} \tag{3.29}$$

with two T-dependent parameters b and σ. In a Gaussian distribution, b would be the maximum position and $\sigma/\sqrt{2}$ the standard deviation; for the skewed Gaussian these identities do not hold, but the parameters are still related to position and width of the distribution.

Contact probabilities

Another quantity accessible using FRET and especially TTET is the probability P_{cnt} for the chromophore contact to be closed. In SAMC, it can either be obtained from $p_T(d)$ as

$$P_{\mathrm{cnt}}(T) = \sum_{d=d_{ij}^{\mathrm{HS}}}^{d_{ij}^{\mathrm{SW}}} p_T(d) \tag{3.30}$$

or directly by counting the number of contacts during simulation. Similarly, the contact probabilities of every individual pair of side chains or of backbone hydrogen bond partners are accessible in simulation, producing a $N \times N$ matrix. The sum over such an H-Bond matrix equals the total number of H-Bonds n_{HB}, and a summation over a selection of cells yields the total number of certain types of contacts – e.g. all α-helical H-Bonds, defined as contacts between NH_i and CO_{i-4}.

Chapter 4

Short peptides

The results obtained with the model and method described on the previous pages will be divided into three chapters. In the present chapter, the available tools of data analysis will be introduced using PRIME20 simulation results for polyserine (polyS), polyalanine (polyA) and polyglutamine (polyQ) chains. The number of repeat units N in these models varies between 6 and 23, which allows a discussion of structure formation depending on temperature, chain length and amino acid type. The second results chapter is a comparison of the three model variants PRIME20, PRIME20s and PRIME20n, mostly using S_{16} and Q_{16} as examples. Finally, the third chapter refers to the experimental FRET/TTET set-up, investigating the influence of chromophores and of a tail added to enhance solubility on structure formation and comparing some results to observations from the experiments.

The variants PRIME20, PRIME20s and PRIME20n, as introduced in section 2.4, differ in their use of squeeze factors. PRIME20n is – barring minor details – identical to the model as it is being used by its creators, the C. K. Hall lab at North Carolina State University. As will be seen in chapter 5, the structures formed by PRIME20n are much more in accordance with observations from literature than those in PRIME20 or PRIME20s. It would therefore be intuitive to use PRIME20n in chapters 4 and 6 as well. However, because the discovery of the backbone squeeze factors came at a rather late stage of the project, the larger-scale N-dependent simulations presented in the current chapter could not be repeated. In addition, many of the observations from PRIME20 are likely to remain valid in PRIME20n.

Regarding the main effects which will be presented, firstly, the formation of helices by polyS and polyA is seen in both PRIME20 and PRIME20n and the energies attained in both models are indistinguishable. Secondly, polyQ does not tend to form such helices in either model and instead prefers hairpin configurations, although the preference for single or double hairpins at very low temperatures divides the models. Thirdly, the side chain interaction energies, from which the results in chapter 6 originate, are the same in both PRIME20 variants, so the effects observed here should be similar in PRIME20n. Following these considerations, the use of PRIME20 is justifiable in spite of its shortcomings. The difference between PRIME20 and PRIME20n will be noted wherever it is relevant to the discussion.

The present chapter serves to explain the method of analysis based on microcanonical and canonical thermodynamics and structural observables. The first section leads from the density of states, the primary SAMC result, to an N-dependent diagram of states, which will subsequently be interpreted by analysing the structures at different temperatures. PolyS and polyA both serve as example systems here because their behaviour is similar, so treating them individually would be redundant and their distinctions are best understood by direct comparison. The third and fourth sections treat polyQ in a somewhat condensed manner, but using the same methods. Some of the polyS results are based on simulations performed by Paul Käthner in the framework of his Bachelor's thesis [96].

4.1 Polyalanine and polyserine – thermodynamics

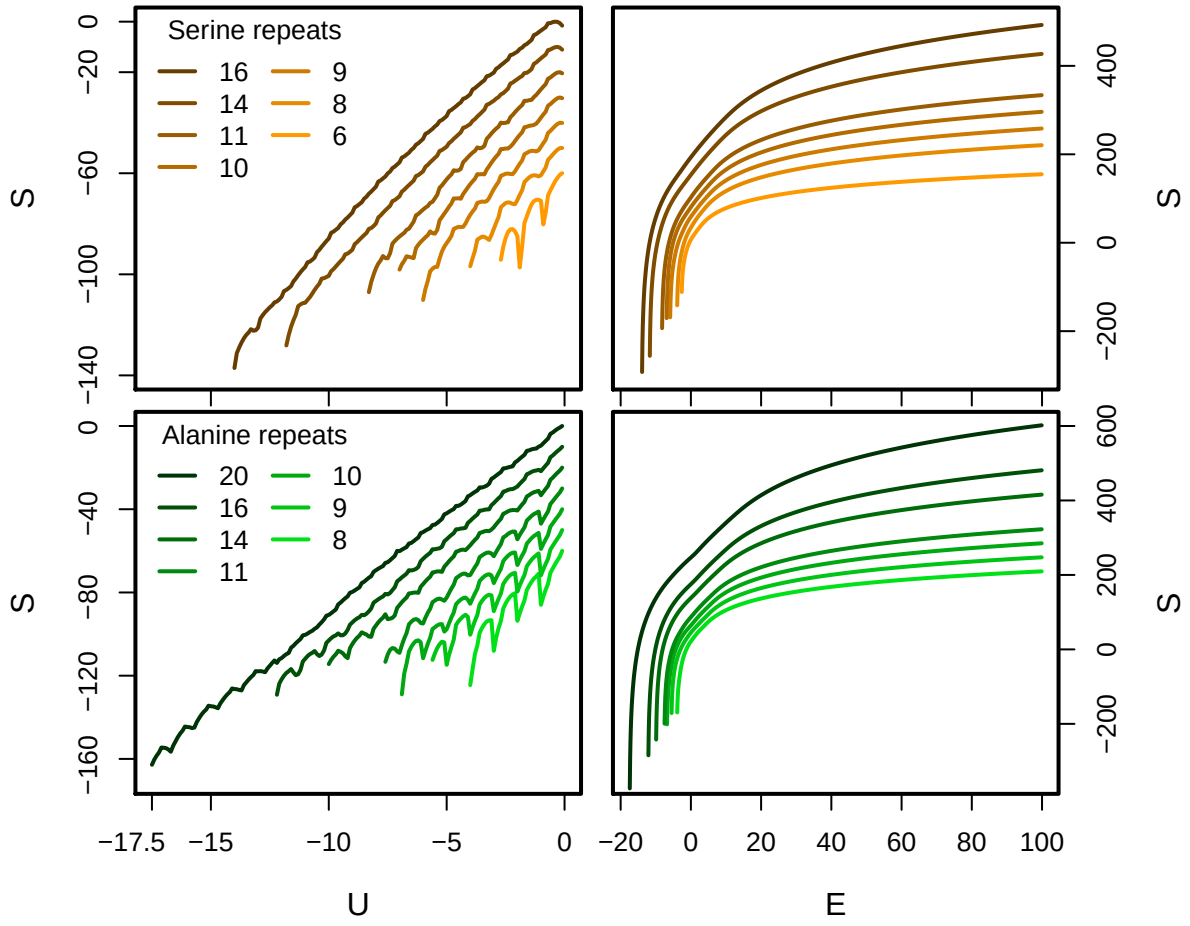

Figure 4.1 – Entropies of polyserine (top) and polyalanine (bottom) in the configurational (left) and full microcanonical (right) ensembles. The graphs are shifted vertically so as not to intersect.

The primary result of an SAMC simulation is the density of states $g(U)$ or its logarithm, the configurational entropy $S(U)$ (eq. (3.21)). As illustrated in the methods chapter, it is often useful to transform these and further quantities to the corresponding ones in the full microcanonical ensemble, $g(E)$ or $S(E)$, according to equation (3.15). Figure 4.1 shows $S(U)$ (left) and $S(E)$ (right) for polyS (top) and polyA (bottom) of all lengths used in simulation. Light colours correspond to short chains.

The shape of these graphs is unusual in comparison to established $S(U)$ graphs [19, 192, 222]. The overall "visual" slope does not change much with energy, which in itself is unexpected, but more importantly the values oscillate in a somewhat regular manner. The oscillations are most pronounced for short chains and can cause large entropy jumps between neighbouring energies. Figures 3.2 A and 4.3 show how these jumps translate to huge peaks in $T(U)$, in positive and negative direction. Negative microcanonical temperatures are not uncommon in simulations of small systems and their physical significance has been discussed at length elsewhere [78, 179], but oscillations like these are unusual and an interpretation of the wild swings in $T(U)$ is not possible.

Due to these oscillations, microcanonical thermodynamic analysis can only take place in the full microcanonical ensemble. Yet, the configurational ensemble is interesting to consider because every structure corresponds to a single energy contrary to the full microcanonical ensemble, where each energy E is a weighted average of all U, hence the mapping of configurations to energies is not unique. The

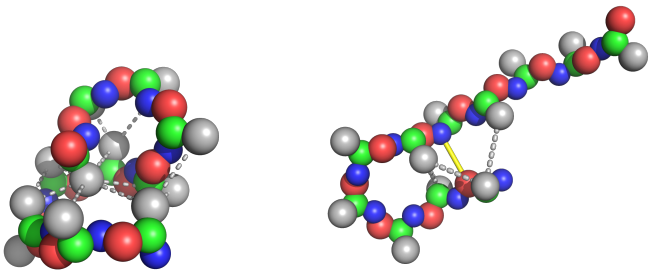

Figure 4.2 – Two A_{11} configuration snapshots from simulation with $U = -1.008$ due to 12 side chain contacts (left) and $U = -1.252$ due to 1 H-Bond and 3 side chain contacts (right), illustrating an increased freedom of movement despite the lower energy. All beads are shown as spheres with 1/4 of their diameter, NH in blue, C_α in green, CO in red and R in grey. The H-Bond is depicted as a yellow stick and all side chain contacts as grey dashed lines.

observed oscillations for example can easily be assigned to a property of the model's energy scales which would remain unnoticed in the full microcanonical ensemble: a homopolymeric system like polyS or polyA has only two interactions with finite energies, the backbone H-Bond and the side-chain contact. The side chain interaction energy is $\epsilon_{\text{Ser-Ser}} = -0.086$ for polyS and $\epsilon_{\text{Ala-Ala}} = -0.084$ for polyA while the H-Bond energy is $\epsilon_{\text{HB}} = -1$. An energy $U \lesssim -1$ can be reached either by closing one H-Bond or through twelve side chain contacts. Because finite positive energies do not occur in these peptides, an energy slightly above $U = -1$ is only reached by closing eleven side chain contacts, which, especially for short chains, is far more restrictive on the overall configuration than a single H-Bond (see fig. 4.2 for examples of both situations, with several entirely free residues in the right-hand structure). Thus the configurational entropy at $U \gtrsim -1$ is lower than at $U \lesssim -1$. The situation is similar in the vicinity of each integer potential energy value, ultimately leading to the observed regular oscillations.

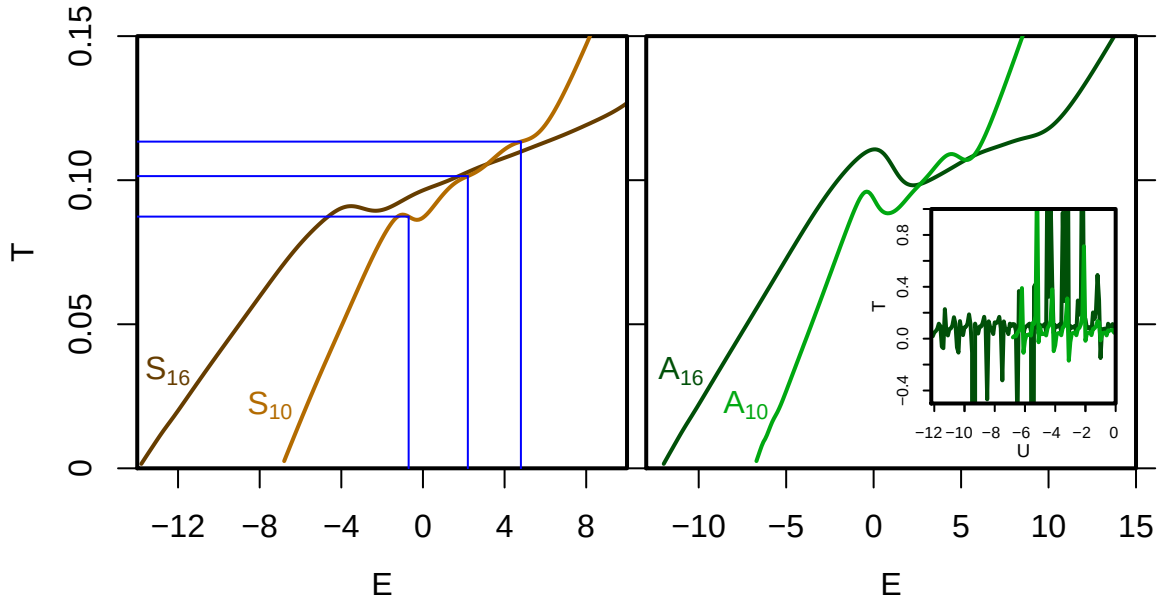

Figure 4.3 – Microcanonical temperatures $T(E)$ of S_{10}, S_{16}, A_{10} and A_{16}. The blue lines mark inflection points in the S_{10} graph. The inset shows the temperature in the configurational ensemble $T(U)$ of both polyA chains.

While the oscillations in the $S(U)$ graphs quickly catch the attention of the viewer and can be interpreted quite easily, the much smoother $S(E)$ curves (right side of fig. 4.1) do not reveal their contained information as readily. The graphs are monotonous and smooth, their slopes decrease steadily with the exception of a few bends around $E \approx 0$, whose positions cannot be located visually. In order to extract

the thermodynamic information, it is necessary to turn to derivatives of $S(E)$. The inverse of the first derivative is the temperature $T(E)$ (eq. (3.22)), shown in figure 4.3 for the peptides S_{10}, S_{16} (left), A_{10} and A_{16} (right). The graphs have an overall ascending shape, intuitively connecting high temperatures to high energies. Inflection points act as indicators for transitions between different states (corresponding to phase transitions in infinite systems), and if the slope of $T(E)$ at an inflection point is negative (called a "Gibbs loop"), the transition has first-order character. In these cases, two states coexist at the transition temperature and their energies are equal to the boundaries of the Gibbs loop.

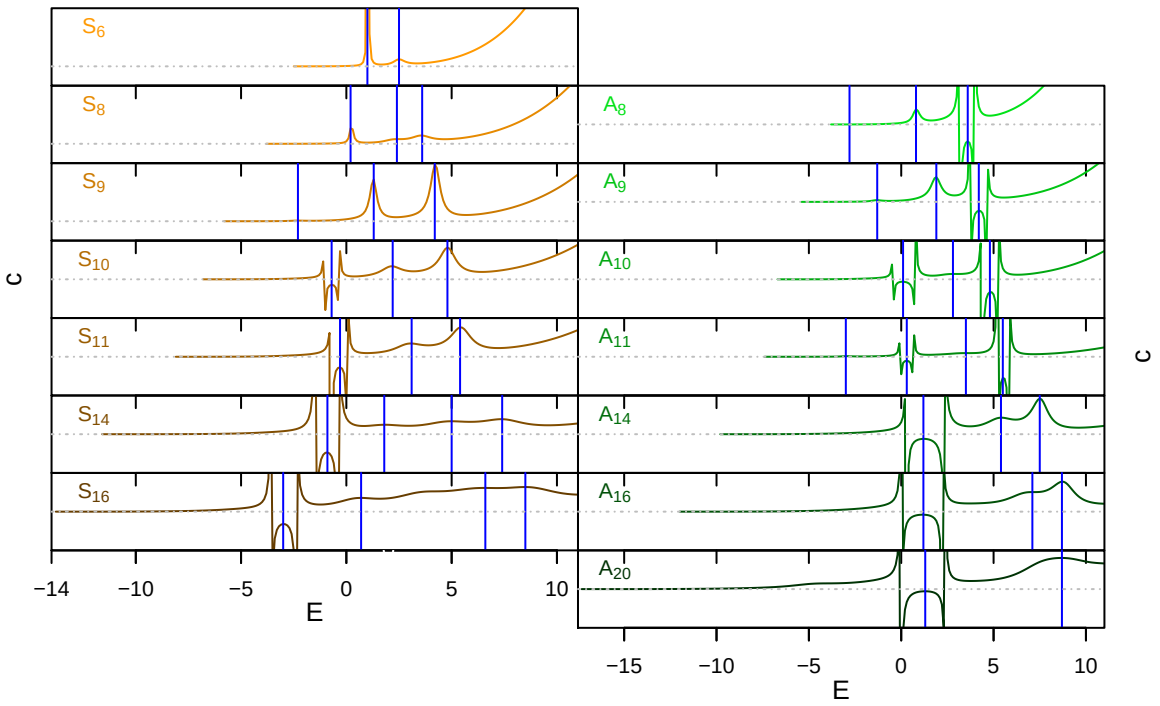

Figure 4.4 – Microcanonical heat capacities of polyS and polyA. Vertical blue lines indicate local maxima (see also fig. 4.3). A maximum with a negative function value neighboured by two singularities indicates a first-order transition. The vertical axes are not labelled, but $c = 0$ is marked by a grey dashed line in each graph.

The S_{10} graph has three inflection points, marked by blue lines. They have been identified by finding maxima of the heat capacity $c(E)$, which is essentially the second derivative of $S(E)$ (eq. (3.23)). A Gibbs loop in $T(E)$ becomes a structure consisting of a local, negative maximum neighboured by two singularities in $c(E)$ and a simple $T(E)$ inflection point becomes a positive local maximum. These structures can be seen in figure 4.4 for all polyS and polyA systems. The vertical scales are unlabeled because the accurate values are not relevant to the analysis; only the positions of local maxima (marked by blue lines) and the information whether they are above or below zero (grey dashed axis) are required.

Considering S_{10} as an example, $c(E)$ has three maxima at $E = \{-0.7,\ 2.2,\ 4.8\}$. The two maxima at positive energies are above zero and indicate second-order transitions. The $E = -0.7$ maximum is negative and has the two neighbouring singularities[1] typical for a first-order transition. The $T(E)$ graph in fig. 4.3 yields three temperatures $T^* = \{0.087,\ 0.101,\ 0.113\}$ corresponding to these maxima. Transition temperatures for all chains can be obtained in the same way. The respective energies are marked in the $c(E)$ graphs and the temperatures T^* are listed in table 4.1. The table is grouped by visual trends in fig. 4.4, even though at this point of the analysis the respective physical significance of the signatures is still unclear, making the grouping a somewhat arbitrary choice.

[1]The "singularities" are sharp peaks instead due to the small system size and energy binning.

Table 4.1 – Transition temperatures of polyS and polyA. Temperatures are grouped vertically by visual trends in fig. 4.4 and first-order transitions are marked red.

polyS					polyA				
S_6		0.087		0.102					
S_8		0.076	0.097	0.105	A_8		0.019	0.081	0.105
S_9	0.057		0.095	0.112	A_9		0.070	0.100	0.107
S_{10}	0.087		0.101	0.113	A_{10}		0.093	0.100	0.108
S_{11}	0.089		0.106	0.115	A_{11}	0.063	0.091	0.104	0.111
S_{14}	0.096	0.100	0.111	0.119	A_{14}		0.105	0.109	0.114
S_{16}	0.090	0.098	0.115	0.121	A_{16}		0.104	0.112	0.115
					A_{20}		0.106		0.113

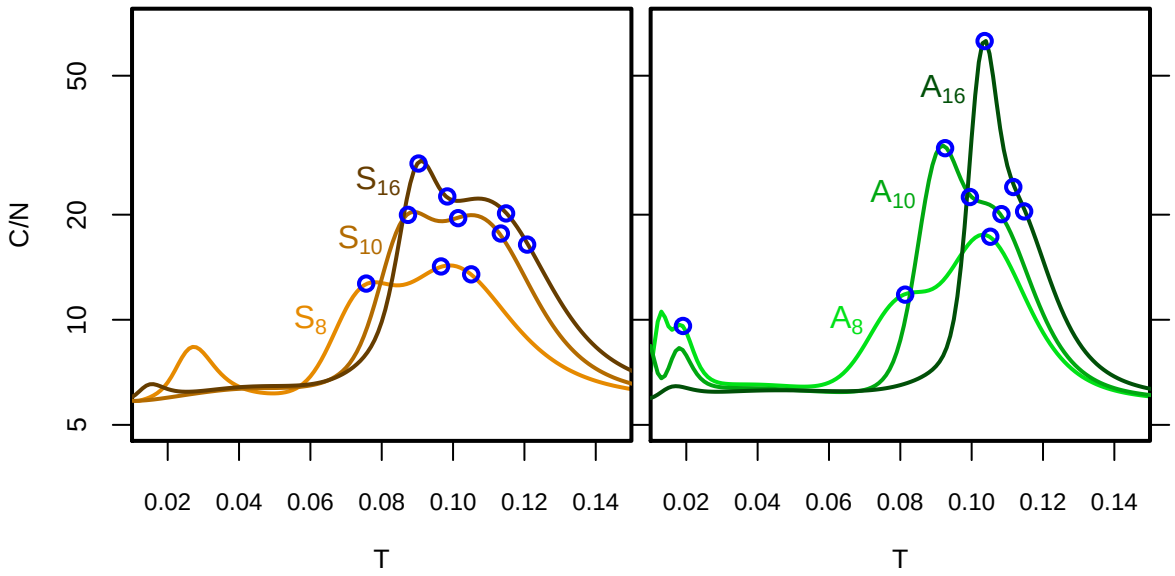

Figure 4.5 – Canonical specific heat capacities of $S_{8,10,16}$ and $A_{8,10,16}$. Microcanonical transition temperatures identified using $c(E)$ and $T(E)$ (see also figs. 4.3 and 4.4) are marked by blue circles.

Before a discussion of the values in the table and the corresponding diagram of states, the microcanonical transition temperatures should be confirmed by regarding the canonical ensemble. Since the majority of experiments – and accordingly most simulations – take place in a temperature-controlled environment, the canonical ensemble is the obvious choice for analysis. However, as figure 4.5 illustrates, the microcanonical analysis tends to provide more detailed results, in this case several transition signatures which are missing in the canonical ensemble – these are relevant data points for the sake of a complete picture, although many of the additional signatures will turn out not to be interpretable physically.

The figure shows canonical heat capacities of three polyS and polyA chains (lengths 8, 10 and 16) divided by chain length. The corresponding transition temperatures from the microcanonical ensemble are marked by blue circles. The vertical axis is logarithmic, but like in the $c(E)$ case before, the exact heat capacity values are not of great interest. The only somewhat curious feature involving these values are the generally higher and narrower peaks in the polyA graphs compared to polyS. They are caused by the smaller difference between the highest and lowest polyA transition temperatures, which for a similar energy gain (documented in table 4.2, p. 50) means higher function values as $C(T)$ is the derivative of the internal energy.

Notwithstanding this detail, the graphs all have somewhat similar features, consisting of two maxima somewhere around $T \approx 0.1$ and in some cases further signatures at very low T (< 0.03). Maxima in $C(T)$ correspond to state transitions and sharp peaks indicate first-order character, although the definition of "sharpness" is much more vague than that of the microcanonical first-order signatures.

The low-T signatures will mostly be ignored; in many cases they do not have a microcanonical counterpart, indicating that they just arise by amplification of insignificant effects due to insufficient convergence or statistics. Some of signatures which are also found in the microcanonical ensemble will be discussed because the corresponding configuration changes are interesting, but even then it should be kept in mind that the temperatures are far too low to be of any relevance in nature. Thus, most of the analysis will focus on the transitions at $T \approx 0.1$, which is closer to physiological temperatures. (A more thorough discussion of physical temperature scales, i.e. of the correspondence between the reduced scale and a Kelvin scale, takes place in section 6.1.)

These peaks at higher T lend themselves to further analysis. As a somewhat intuitive expectation, the peak at higher temperature may be attributed to a collapse driven by interaction of the side chains (often called a hydrophobic collapse[2]) and the lower peak to a transition from a disordered globule to a regular, possibly native, folded structure dominated by backbone hydrogen bonds. This intuitive assignment will be confirmed in the next section where conformational observables are discussed.

An interesting effect in the $C(T)$ graphs is the change of relative peak heights: for $N = 8$ the collapse peak is higher while for $N = 16$ the folding peak dominates. As will be seen in the next section as well, a regular helical structure of length 16 has a clearly lower energy than a globule or hairpin, but this difference between helix and globule is much smaller for $N = 8$. The peak heights, correlated to the energy change during a transition, reflect this property.

Finally, the position of the blue circles, i.e. of the microcanonical transition signatures, needs to be discussed. In all cases, one of the signatures is located close to the folding maximum of $C(T)$ and can therefore be assumed to describe the same transition. As figure 4.4 and table 4.1 show, the transition has first-order character for all $N \geq 10$. On the other hand, the collapse maximum of $C(T)$ usually coincides with two microcanonical signatures (three for S_{16} and S_{14}, only one for A_8). This indicates the existence of another state, which is resolved in the microcanonical, but not in the canonical ensemble. However, a clear distinction between this and the globule state based on conformational observables does not exist and since the canonical transitions are not distinguishable either, it does not seem useful to pay attention to these additional microcanonical transition signatures as long as no structural change is found at these energies. Thus, the analysis will be focussed on the highest-temperature signature, interpreting this as the temperature of collapse from a random coil into a random globule, and the folding signature, which is well-defined in the majority of cases.

The above discussions ultimately lead to the diagram of states shown in figure 4.6. Here, the transition temperatures T^* of polyS and polyA (table 4.1) are drawn as functions of the chain length N. First-order transitions are marked by crosses, and the points which are attributed to the coil-globule and globule-native transitions are connected by lines, forming boundaries between the three states: the random coil at high T, the native state at low T and the globule state in between. Even though the deviations of single points from the general trends suggest a substantial uncertainty in T^*, it becomes apparent that the polyA signatures lie closer to each other than those of polyS, narrowing the temperature range of the globule state. Short polyalanines are often discussed to be two-state folders with a single coil-helix transition. In this model, polyA does not quite fold in this manner, but with its narrow globule state it might be called more "two-state-like" than polyS. Due to the shortcomings of PRIME20 noted above, a more accurate discussion is not of great use, especially since in PRIME20n both peptides are closer to two-state behaviour, as will be seen in chapter 5.

[2]The serine side chain is in fact polar. The difference to the hydrophobic alanine cannot be expressed by an implicit-solvent homopolymer simulation, but the mapping to physical temperatures in section 6.1 will be vastly different for this reason.

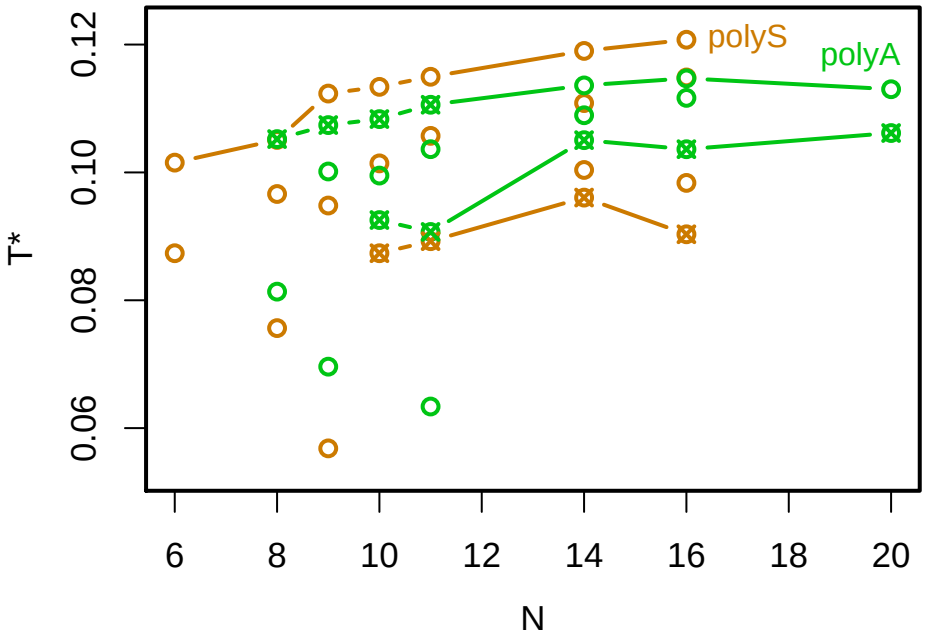

Figure 4.6 – Diagram of states for polyserine and polyalanine. Transition temperatures taken from the microcanonical heat capacity maxima are shown as empty circles, 1st-order transition signatures as crossed circles. Lines connect the points which will be referred to as transition temperatures for the remainder of the chapter. They were obtained by a comparison between the microcanonical and canonical ensembles.

4.2 Polyalanine and polyserine – configurations

The diagram of states (fig. 4.6) is based solely on thermodynamic quantities. It was speculated to show transitions between random coil, globule and native states with the only justification being expectations from literature or previous knowledge. The next essential step is the analysis of conformational quantities to confirm the assignment of states and to gain insight into their proterties, especially to identify the character of the native state.

A commonly employed conformational quantity, especially in polymer science, is the squared radius of gyration, R_g^2 (eq. (3.25)). It maps a complex molecular structure to a one-dimensional variable which is easier to understand, but of course does so at the expense of potentially important details. This makes R_g^2 a useful starting point of analysis, although it will not provide a complete picture of the structures.

The squared radii of gyration of S_{10}, S_{16} and A_{16} are depicted in figure 4.7. Subfigures A and B show the canonical average $R_g^2(T)$ and its derivative with respect to temperature, subfigures C and D the microcanonical and configurational ensemble averages. Blue circles in subfigure B mark the transition temperatures from table 4.1.

Both the canonical and microcanonical graphs are dominated by a large increase of R_g^2 around the transition temperatures or energies, supporting the hypothesis of a transition between an extended random coil at high temperatures (energies) and a compact globular structure at lower T (E). The polyA random coil is slightly more extended than that of polyS (consistently for all N, seen here for $N = 16$), indicating a lower flexibility of polyA in the model. In both cases, the values generally increase with N.

Because the coil-globule transition dominates the shape of the graphs, further transitions are not well identifiable using R_g^2. The derivatives (subfig. B) have a double peak structure similar to $C(T)$ (fig. 4.5), but if a characteristic radius of gyration for an intermediate state was to be defined, a plateau at this state would be required, which does not occur in any of the $R_g^2(T)$ graphs. Like the random coil, the

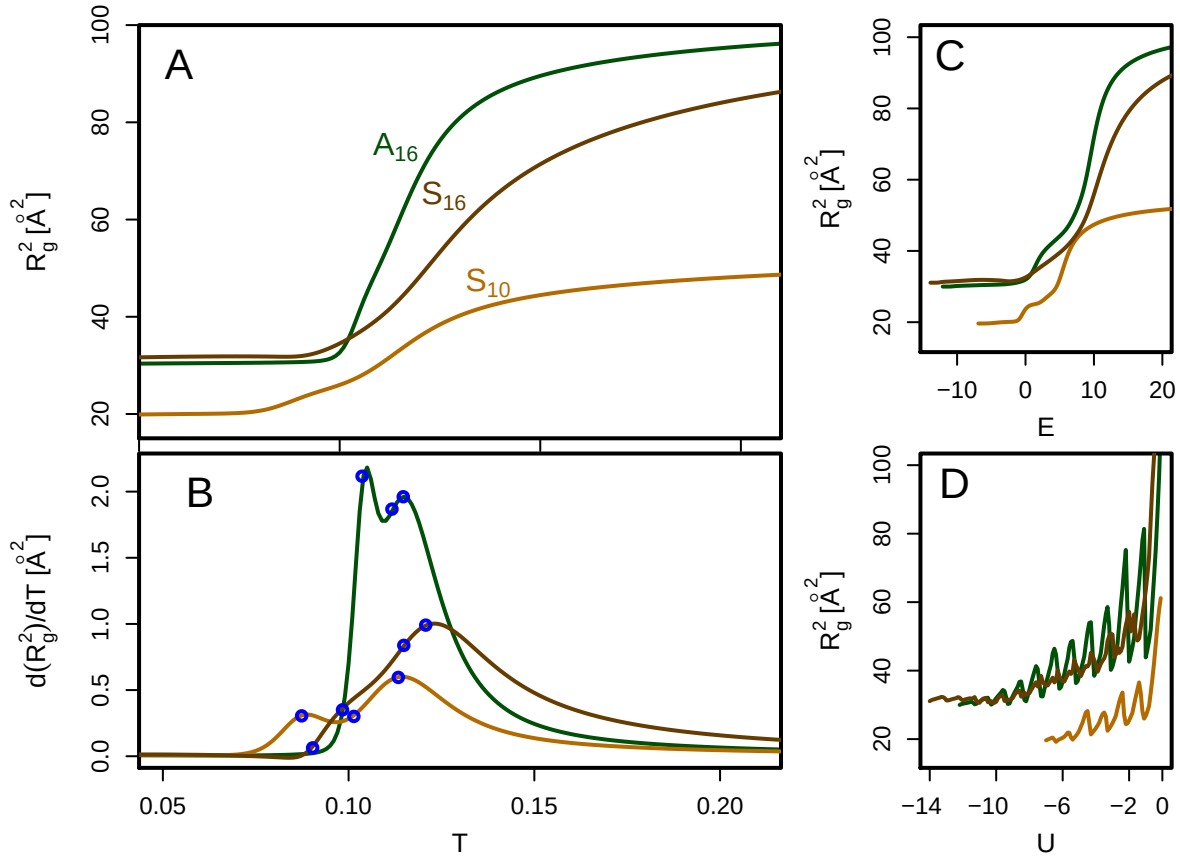

Figure 4.7 – Squared radius of gyration in the canonical (A), microcanonical (C) and configurational (D) ensembles and its canonical derivative (B) for the three peptides S_{10}, S_{16} and A_{16}.

globule state evolves with changing temperature or energy. At low T or E, all graphs stabilise at a well-defined value (20 Å^2 for S_{10}, around 30 Å^2 for S_{16} and A_{16}), indicating a well-defined, unchanging native structure, but only S_{10} has an intermediate state with $R_g^2 \approx 26$ Å^2 which is stable over several consecutive energy bins.

Lastly, the $R_g^2(U)$ graphs are worth a short consideration. Exemplary for all observables, this quantity oscillates in a manner similar to $S(U)$ (fig. 4.1). As discussed there, the oscillation is caused by the difference in energy scales between H-Bonds and side chain interactions. With decreasing potential energy, a chain alternates between coil-like and globule-like conformations, leading to large variations in R_g^2 until at low energy the high number of H-Bond interactions does not allow an extended coil state any more and the oscillations abate. The significance of this is that not only the entropies fluctuate a bit, but in fact the whole system alternates between different states, which is the fundamental reason why analysis in the conformational ensemble is so impractical for this model. The full microcanonical and canonical systems, in a manner of speaking, "sort" these intertwined states on their E or T axes.

Another reduction of complex configurations to a low-dimensional quantity is presented in figure 4.8. The four colour maps show the probability $p(d, T)$ to find the terminal side chain beads, i.e. R_1 and R_N, at a certain distance d (or, more accurately, in the interval between d and $d + 0.1$ Å) at temperature T. The colour scale ranges from white via turquoise to red as defined in the figure caption. Transition temperatures are marked by vertical blue lines. The two figures in the rightmost column show the most probable distance versus N at specified temperatures. It should be noted that d is not identical to the end-to-end distance R_{EE}, which is measured between NH_1 and CO_N, but due to the restricted chain geometry both quantities behave similarly.

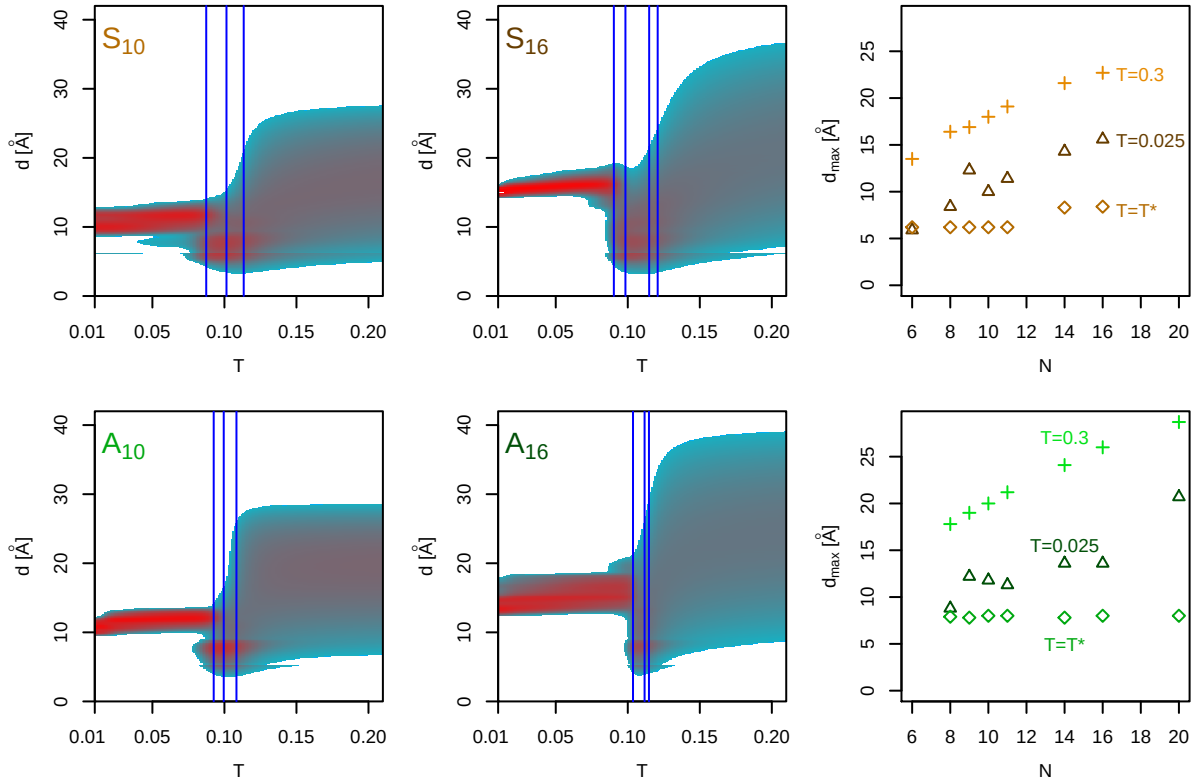

Figure 4.8 – Probability $p(d, T)$ for the terminal side chains to be found at a distance d at temperature T. Red colour indicates high probability (up to 0.1 in a bin of width 0.1 Å), white lower probability (below 0.001 in such a bin). The graphs in the rightmost column show the positions of the probability maxima versus chain length at three distinct temperatures $T = 0.3$ (above all transitions), $T = 0.025$ (below all transitions) and $T = T^*$ in between, which of course varies with N.

In these plots, the difference between the three states is more visible than in $R_g^2(T)$. Each of the four images consists of three distinct sections: the first at temperatures below all transitions, features a narrow distribution around 10 Å for $N = 10$ or 15 Å for $N = 16$. The second section, between the lowest and highest transition temperature, is dominated by a maximum around 8 Å independently of N, and in the third section at high temperatures, the probability distribution is very broad and its maximum lies at higher d, roughly 20-25 Å.

The assignment of the three states is supported by these observations. In a random coil with high energy, i.e. a low number of attractive interactions, the chain behaviour is dominated by geometric constraints and repulsions, which according to generic polymer models leads to a kind of bell distribution like the one seen at high temperature. A more accurate description follows in the discussion of fig. 4.9.

As the chain collapses into a denser globule at intermediary temperatures, the distribution becomes narrower and the preferred distance is lower. A noteworthy feature here is the jump at 6.4 Å in case of polyS and 5.4 Å in case of polyA, visible as a horizontal "fault line" in the colour maps. These two distances are the respective square well diameters, which means that the jump is caused by the shape of the potential. A configuration in which the two beads are just in contact has a lower potential energy than a configuration in which they are just out of contact, but the conformational freedom and therefore entropy is almost identical in both cases. Hence, the in-contact case will always have a lower free energy than the out-of-contact case, leading to a preference seen in the form of this jump. The apparently bivariate shape of $p_T(d)$ in the globule state is therefore not an indicator of two coexisting states, but rather a symptom of the modelling choices.

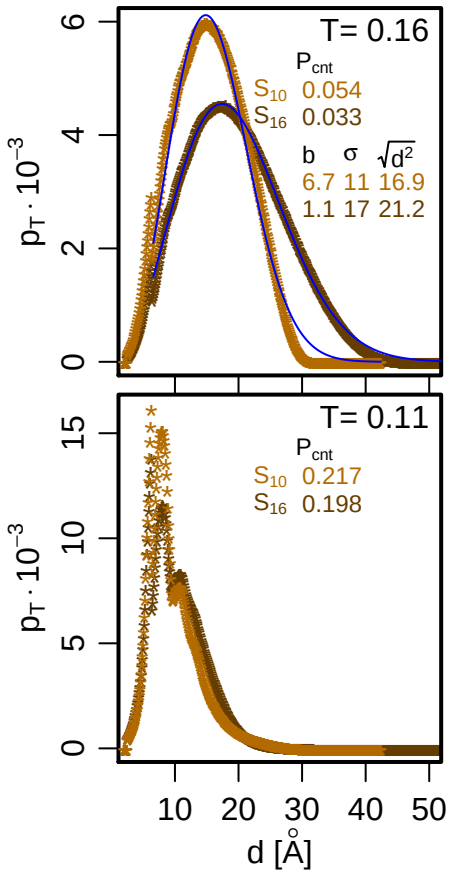

Figure 4.9 – Distance probabilities $p_T(d)$ of polyserine at $T = 0.16$ and 0.11. The graphs represent vertical cuts through the 2D images in fig. 4.8. At the higher temperature, a distribution function from the Edwards model fits the curve above the contact distance $d^{SW}_{\mathrm{Ser,Ser}}$ well, at the lower temperature the fit does not converge.

Finally, at lowest temperatures, the maximum of $p_T(d)$ is found at a higher terminal side chain distance. This means that the folded configuration is in a sense more extended than the globule, although the radius of gyration (fig. 4.7) is lower. With the well-documented common structure motifs of helices and hairpins found in polymers and especially proteins in mind, the separation of termini hints at the folded state being a helix opposed to the globule state which might have hairpin properties. A similar duality between an extended helix and a globule state with shorter end-to-end distance has been reported for polyA in the past as well [44].

The rightmost column of fig. 4.8 shows the dependence of the maximum positions on N discussed in the previous paragraph. The selected temperatures are $T = 0.3$ in the random coil region, $T = 0.025$ in the folded region and $T = T^*$, which is the collapse temperature identified earlier. (Note that T^* varies with N.) Because the conformational collapse generally happens at slightly higher temperatures than the energetic collapse (e.g. the maxima of conformational observables like $\partial\langle R_g^2\rangle/\partial T$ lie at higher T than those of $C(T)$), this temperature already lies inside the globule region.

As discussed above, the distances are highest in the random coil state and lowest for the globule. An interesting observation here is the trend versus N: in the coil and folded states, the distance increases – apparently linarly – with N, another hint at the folded state being a helix whose length increases with each added monomer. In contrast, the maximum position in the globule state is independent of N, except for S_{14} and S_{16}: the maxima of the shorter polyS chains lie at 6.3 Å, just below the square well diameter, and those of S_{14} and S_{16} at 8.5 Å, closer to the values seen in polyA (7.9-8.1 Å). As discussed before, both distances belong to the same, seemingly bivariate, distribution, so the preference for the lower or higher distance is largely arbitrary and bears no physical significance.

A vertical cut through the 2D images in figure 4.8 yields distance distributions at fixed temperatures, shown in fig. 4.9 for S_{10} and S_{16} at $T = 0.16$ (coil state) and $T = 0.11$ (globule state). The high-temperature distribution is broad and bell-shaped as mentioned before, except for the documented jump at $d = 6.4$ Å. The blue lines overlaying the bell part are fit curves according to the Edwards model [59] for a generic polymer with excluded volume. The function which results from the Edwards model is a skewed Gaussian (see eq. (3.29)) with two free parameters, b and σ, controlling the maximum position and distribution breadth. Despite the complex polypeptide geometry, the simple model fits the simulated data reasonably well at this high temperature. The resulting parameters b and σ (in Å) are listed in the figure, but without comparison to further simulations or experiments they are not of great use. A comparison to experimental results takes place in chapter 6. Another result obtained from the fit is the root mean square distance between the terminal side chains, $\sqrt{\langle d^2\rangle} = \left(\int_{d_{\min}}^{d_{\max}} d^2 p(d)\mathrm{d}d\right)^{1/2}$. Similarly to the radius of gyration, it is a measure for the extension of the chain and therefore intuitively larger for

S_{16} than for S_{10}. Lastly, another set of numbers given in the figure is the probability of a contact, i.e., the integral of $p(d)$ over the distances smaller than d_{ij}^{SW}. It can be calculated from the fit function or – the option chosen here – directly from the data.

The fit ignores the lowest distances at which the side chains are in contact because their attraction is not included in the Edwards model. Integrating over this low-distance region yields the contact probability P_{cnt}, which is printed in the plot window for both chains. At $T = 0.16$, these probabilities are rather low, about 5% and 3% for S_{10} and S_{16} respectively, compared to the values in the globule state (bottom graph of fig. 4.9). In this state, the terminal side chains have a probability of 20% to be in contact, of course matching the overall denser configuration seen in a qualitative way in fig. 4.8. Meanwhile, b and σ cannot be compared between the different temperatures because the fit does not converge at $T = 0.11$ or below. The out-of-contact distribution is much narrower and has two maxima which cannot be represented by the simple function from the Edwards model. While the random coil state is defined by a low overall number of interactions and therefore well described by this non-interacting model, the globule state features a higher number of contacts, especially between side chains, and thus cannot be described by the model any more. In this sense, the convergence of the fit is another indicator of the collapse transition.

The above discussion of $p_T(d)$ provides only little insight regarding polyS and polyA themselves here, but it has turned out to be valuable for comparison between simulation and spectroscopy experiments, which are discussed in more detail in chapter 6. For the present chapter this overview of available parameters shall suffice. In order to identify the chain configurations, especially in the folded state, accurately, a different representation of results is needed.

Such a representation is shown in figure 4.10. It contains six contact maps (or, synonymously, contact matrices) averaging the conformations of certain chains at certain temperatures. Subfigures A-C are the Backbone H-Bond contact matrices (short: HB matrices) of S_{16} at $T = \{0.16,\ 0.12,\ 0.06\}$ corresponding to the coil, globule (just below the transition) and folded states. For comparison, subfigure D shows the folded state of S_{10} and F that of A_{16}. Subfigure E is a side chain contact map of S_{16} at $T = 0.06$. The format is the same in all matrices: every cell represents one possible contact (i, j), with $(1, 1)$ lying in the top left corner as outlined by the axis labels, colours encode the contact probability by intensity ranging from white (0) to black (1). The base colours have no further meaning, they were selected to comply with the general colour scheme used to distinguish polyS and polyA throughout the chapter. To enhance contrast, the intensities are scaled up 100-fold in subfigure A, meaning that black represents a probability of 0.01 instead, and 10-fold in subfigure B.

Supplementary to the contact maps, figure 4.11 depicts configurations of S_{16} at three temperatures and of A_{16} at low T, matching subfigures A, B, C and F as labelled. Individual contact maps of these configurations are attached. (Note that the "probabilities" in this case can only be 0 or 1.) These images are singular snapshots from simulation. The NH, C_α and CO beads are represented as blue, green and red beads with 1/4 of their actual diameters, side chains are not shown and yellow sticks indicate the presence of H-Bonds. (In the case of A_{16}, two H-Bonds are coloured purple to highlight an effect to be discussed below.)

One feature which is common to all HB matrices is a white region comprising the descending main diagonal and three further diagonals to each of its sides. The model does not allow H-Bonds between beads separated by less than three intervening residues, so the closest occupiable fields to the main diagonal are those of $(i, i \pm 4)$ contacts. In figure 4.10 E, only the main diagonal and two closest side diagonals are entirely unoccupied because a side chain bead cannot interact with itself by definition and with its direct neighbours due to the chain geometry, but it can interact with the next neighbours.

Figures 4.10 A-C follow the folding process of S_{16} from high to low temperature. Subfigure A ($T = 0.16$) is characteristic for the random coil state: the average potential energy is high ($U = -0.94$ at this temperature), the number of H-Bonds, i.e. the sum of values in all matrix cells, very low (≈ 0.09). A

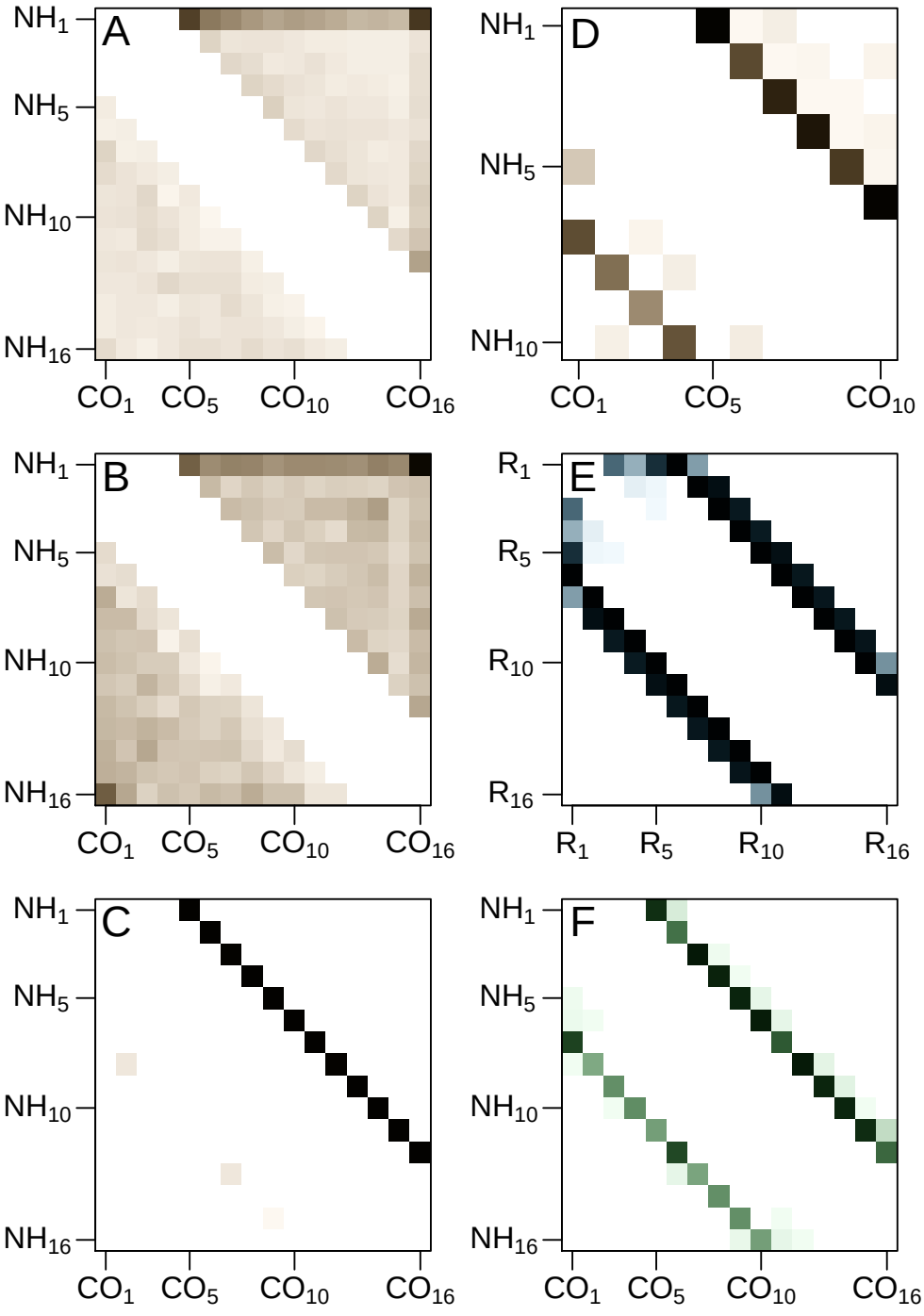

Figure 4.10 – A-C: Backbone hydrogen bond contact matrices of S_{16} at three temperatures above, between and below the $c(E)$ signatures (cf. fig. 4.6). D-F: Low-temperature contact matrices of S_{10} backbone (D), S_{16} side chains (E) and A_{16} backbone (F). Each cell represents one pair of beads. Dark colours indicate high probabilities for the contact to be closed. The values in A and B are scaled up 100-fold and 10-fold respectively.

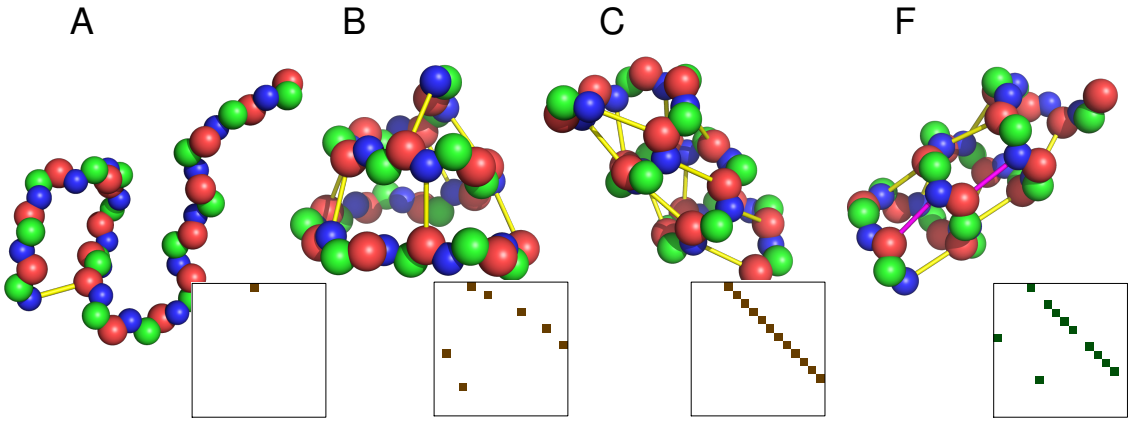

Figure 4.11 – Configuration snapshots of S_{16} at three selected energies and of A_{16}, corresponding to fig. 4.10 A-C and F. All beads are drawn with 1/4 of their actual diameter for clarity of view; side chains not shown. Yellow sticks indicate the positions of H-Bonds, which are also shown in the respective contact matrices. The two magenta sticks in F are H-Bonds which break the $(i, i+4)$ pattern in comparison to C.

preference for H-Bonds involving the terminal beads can be observed because an H-Bond in the interior of the chain requires an overall more compact structure with an accordingly lower energy and is therefore less likely to be formed at such a high temperature. The N-terminus seems to be preferred over the C-terminus too, possibly due to its higher flexibility as the NH bead is smaller than the CO bead. The most probable contact is $NH_1 \cdots CO_{16}$, but it should be noted that even this H-Bond is only formed with a probability of 0.5%. Figure 4.11 A is an exemplary snapshot of such a configuration with an H-Bond between NH_1 and a central CO bead.

Subfigure B at $T = 0.12$ depicts the globule state of S_{16}. The image looks similar to the coil state as no signature of any regular structure can be observed. The preference for terminal H-Bonds is similar to the coil state, but at a much higher level now: the $NH_1 \cdots CO_{16}$ bond has a probability of 9%, and overall an average of 1.8 H-Bonds are closed. Both values increase steeply with reduced temperature. In contrast to subfigure A, there is a notable contribution of the $NH_{16} \cdots CO_1$ contact (4%), which indeed is often formed in combination with $NH_1 \cdots CO_{16}$. Because the energy gain by two H-Bonds is rather large, structures like these are suppressed at $T = 0.16$, but as seen here, they are more common at $T = 0.12$.

Another common combination of H-Bonds is that of $NH_1 \cdots CO_{16}$ and $NH_3 \cdots CO_{14}$. If the chain ends align in an antiparallel fashion, these two contacts are available, but the NH_2 and CO_{15} beads are turned outwards, preventing the corresponding contact in such a conformation. Due to the preference for end-to-end contacts, most H-Bonds involving the residues 2 or 15 are therefore slightly suppressed, as seen in the HB matrix by slightly lighter colours in the second row and penultimate column.

The configuration snapshot in fig. 4.11 B depicts a denser globule. While the average number of H-Bonds at $T = 0.12$ is 1.8, this configuration contains 7 H-Bonds. It is observed most likely at $T \approx 0.095$, still above the folding transition ($T = 0.090$). This illustrates the broad definition of the "globule" state, ranging from dense configurations like this to the more expanded ones seen in the HB matrix.

Figure 4.10 C, showing the folded state at $T = 0.06$, looks entirely different. Most H-Bonds including the formerly prominent end-to-end contacts have vanished to zero probability to the benefit of a descending diagonal line of $NH_i \cdots CO_{i+4}$ contacts spanning the whole chain from $NH_1 \cdots CO_5$ to $NH_{12} \cdots CO_{16}$. All of these contacts are formed with a probability of almost one, meaning that the structure is regular with only little disturbance. Regular peptide structures are usually helices [13, 31, 53] and the snapshot in fig. 4.11 C confirms this low-T state of S_{16} to be a helix as well.

A structure with regular $NH_i \cdots CO_{i+4}$ H-Bonds has been named a γ-helix by Pauling [156], although the name is somewhat ambiguous nowadays as it has been assigned to other rare [43, 128, 165, 188], ill-defined [229] or newly invented [145] helix types in more recent literature. This is due to the fact that

Pauling's γ-helix has never been observed in experiment and is therefore largely forgotten. (To illustrate, a request for the term "gamma helix" on Google Scholar in January 2019 yielded 35 results in contrast to 61500 for "alpha helix".) Finding a γ-helix as the native state is therefore surprising at the very least and a reason for criticism regarding the model. In fact, this observation has motivated the discussion about PRIME20/PRIME20n which will take place in chapter 5. As will be shown there, the γ-helix is a symptom of shortcomings of PRIME20 and the updated variant PRIME20n indeed produces α-helices instead. In the present chapter however, treating PRIME20 only, the γ-helix is without doubt the native state of polyA and polyS and its real-world implications will not be discussed further.

Even though the γ-helix is a clearly defined and ordered structure, some amount of fluctuation remains possible. Figure 4.10 E shows the side chain contacts of S_{16} at the same temperature as fig. 4.10 C, i.e. in the same state. In the helix core, only the contacts R_i–$R_{i\pm5}$ (each with almost 100% probability) and R_i–$R_{i\pm6}$ (between 70% and 80%) are observed, signifying very little freedom for disorder. Close to the chain ends, especially to the N-terminus, contributions of several further contacts are observed. (Interestingly, the R_1–R_6 contact is closed in 99% of the configurations regardless of this freedom of movement.)

Another source for disorder is seen more clearly in subfigure F, the low-T state of A_{16}. Like S_{16}, the $(i, i+4)$ diagonal is strongly populated, indicating a γ-helical state, but the population is not fully consistent throughout all configurations. Instead, another diagonal, $(i, i-6)$, contributes roughly 25% in most cells. The contacts $NH_7 \cdots CO_1$ and $NH_{12} \cdots CO_6$ are most prominent with a probability of 50% each. Figure 4.11 F depicts a configuration which incorporates these two contacts (highlighted by pink colour) at the expense of three regular γ contacts: $NH_2 \cdots CO_6$, $NH_7 \cdots CO_{11}$ and $NH_{12} \cdots CO_{16}$ are missing in the helix. The amino acids participating in the irregular H-Bonds are practically "turned upside-down", acting as a defect in the γ-helix. The total number of H-Bonds is decreased to 11 in this structure, making it energetically unfavourable, but a clean γ-helix of polyalanine has not been observed during simulations at all, suggesting an essential overlap somewhere in the structure which is resolved by including this defect. Furthermore, the defect can occur in any position along the chain, as indicated by the occupancy of the $(i, i-6)$ diagonal, but the spacing between the involved residues is always the same. (For example, the defective contacts in A_{16} can be $NH_8 \cdots CO_2$ and $NH_{13} \cdots CO_7$ instead of $NH_7 \cdots CO_1$ and $NH_{12} \cdots CO_6$.)

The same defect can occur in polyS as well, seen by two very lightly coloured cells in fig. 4.10 C, however these two cells are only populated by 4% of configurations at this temperature and even close to the helix-globule transition they never surpass 10%. Polyserine in PRIME20 folds from the globule state directly to a mostly unperturbed γ-helix while polyalanine folds into a γ-helix with a defect. Since the side chain energies are almost identical, the difference must be caused by bead sizes and positions (see tables 2.4 and 2.5): despite the additional oxygen atom in the serine side chain, its PRIME20 bead is slightly smaller than that of alanine (2.5 Å vs. 2.7 Å), but due to this oxygen, its center of mass is further outside, so bond the between the C_α atom and the side chain bead is longer for serine (1.967 Å) than for alanine (1.600 Å). It is therefore easily imaginable that the polyA helix might be hindered by additional overlaps which do not happen in polyS and which also cause the larger stiffness in the random coil state compared to polyS (judged by $R_g^2(T)$, fig. 4.7).

Finally, figure 4.10 D depicts the folded state of S_{10}. Like A_{16}, it is a γ-helix with $(i, i-6)$ defects which persist even at the lowest reached energies. This is surprising in comparison to S_{16}, in which the defects do not occur. The reason lies in the total number of H-Bonds: S_{16} reaches up to $12 = N - 4$ H-Bonds, which is the maximum possible number in a γ-helix, where A_{16} only reaches $11 = N - 5$ due to the defects. In the case of S_{10} however, the number of H-Bonds is $10 = N - 4$ regardless of the defect because at this chain length only one γ-type H-Bond must be broken to form one $(i, i+6)$ bond where in the A_{16} helix two defect bonds replaced three γ-type H-Bonds. Thus, the defects in the S_{10} (or A_{10}, whose HB matrix is essentially the same) helix are entropically favourable at no energetic cost and the helix can be formed with or without defects at will.

To conclude the present and previous sections, both polyS and polyA of various lengths were found to fold from a random coil at high temperatures through a disordered globule to a γ-helical native state at low T. The polyA helix contains defects which are energetically unfavourable for $N > 11$, but necessary due to a steric repulsion. At short chain lengths, these defects are found in polyS helices as well. The globule temperature range of for polyS is broader than that of polyA and in both cases this state contains further thermodynamic transition signatures. However, a physical interpretation of these transitions was not possible. It may be worth noting that these transitions are not observed if the updated model, PRIME20n, is used (see chapter 5).

4.3 Polyglutamine – thermodynamics

After the rather extensive introduction to the behaviour of polyserine and polyalanine, the following two sections treat polyglutamine (polyQ) in a more condensed manner. Many concepts are the same as above, the major difference being the inability of polyQ to fold into the γ-helical state due to a steric hindrance by the side chains. The native state is found to be a hairpin configuration instead, which is unfavourable for polyS and polyA due to the lower number of H-Bonds and corresponding higher configurational energy. This result matches the observation from literature that polyQ tends to be disordered or form β-sheets, although once again the interpretation has to be taken "with a grain of salt" since the configurations might (and do) look differently in PRIME20n.

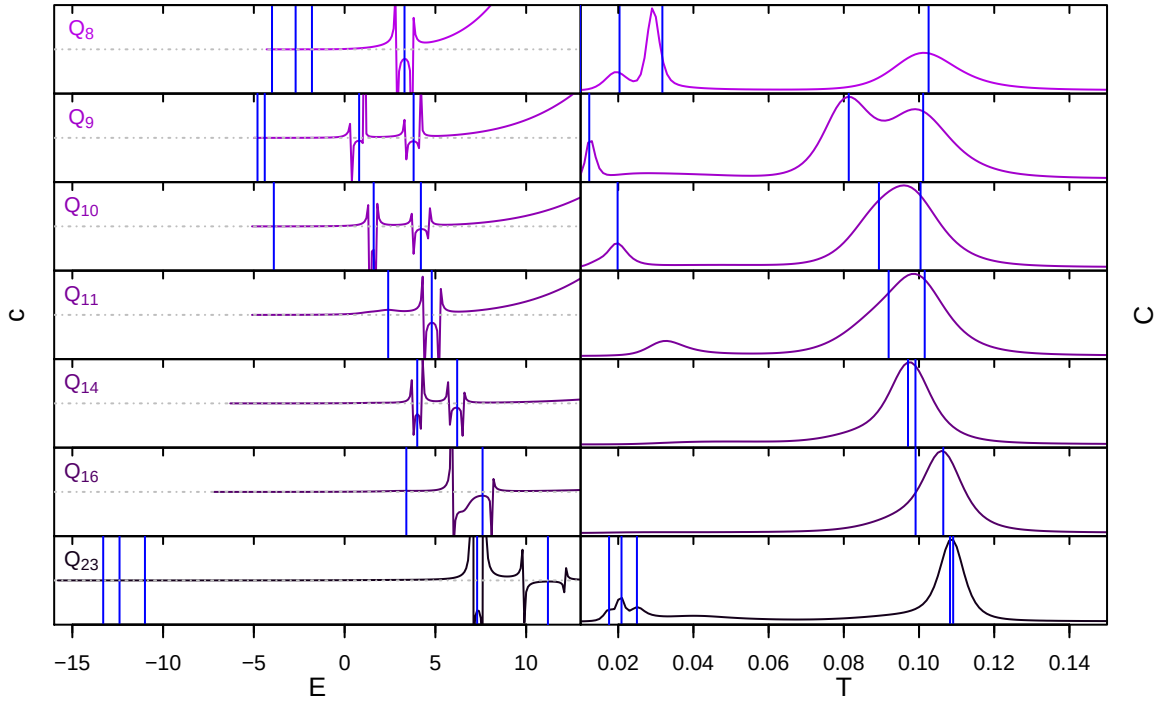

Figure 4.12 – Microcanonical and canonical heat capacities of polyglutamine. Local maxima of $c(E)$ and the respective temperatures in $C(T)$ are marked by blue lines.

As before, transition temperatures are identified by means of the microcanonical and canonical heat capacities $c(E)$ and $C(T)$. Figure 4.12 shows both graphs side-to-side for the investigated systems Q_8, Q_9, Q_{10}, Q_{11}, Q_{14}, Q_{16}, and Q_{23}. The resulting microcanonical transition temperatures and energies are marked in blue.

Unlike polyS and polyA, polyQ has a first-order transition from the random coil to the "globule" state consistently throughout all N, suggesting a higher level of order than in those systems. In many

cases, a second transition with first-order character occurs, the exceptions being Q_8 (without a nearby second signature), Q_{11} (with a second-order signature instead) and Q_{16}. In the latter case, the second transition is visible as a kink in the $c(E)$ graph, but overlaid by the first, whose signature spans a wide energy range. The shape of the graph suggests that the second transition would be of first order too, but its position cannot be identified. In addition to all these transitions, some effects at very low energies and temperatures can be observed for $N \leqslant 10$ and $N = 23$. For short N, these effects are usually considered artefacts of simulation or chain length, but for $N = 23$, they can be assigned to a transition between two distinct configurations, as will be seen at the end of the chapter.

In the canonical picture (right side of figure 4.12), a distinction between the transition signatures is again impossible. With the exception of Q_9, all graphs have a single peak at $T \approx 0.1$, whose position is more or less closely related to both transition temperatures derived from the $c(E)$ maxima. The low-E effects all coincide with $C(T)$ maxima and in case of Q_{11}, another maximum is found at low T which has no microcanonical counterpart. Because the two high-T transitions cannot be distinguished in the canonical picture, most chains can be regarded as two-state folders without a globule state. Only Q_9 may have three distinct states, but the intermediate does not have any defining characteristics (like a plateau in a conformational observable graph) and will not be treated further in this chapter. Contrary to polyS and polyA, the transition temperature from coil to the globule/folded state does not depend strongly on N. The second transition temperature obtained from $c(E)$ increases from $N = 9$ to $N = 14$ where it coincides with the higher temperature and they remain identical for $N = 23$ and probably also $N = 16$ (remembering that the second transition is visible, but not locatable in $c(E)$, so the additional blue line in $C(T)$ is considered to be an unrelated effect).

4.4 Polyglutamine – configurations

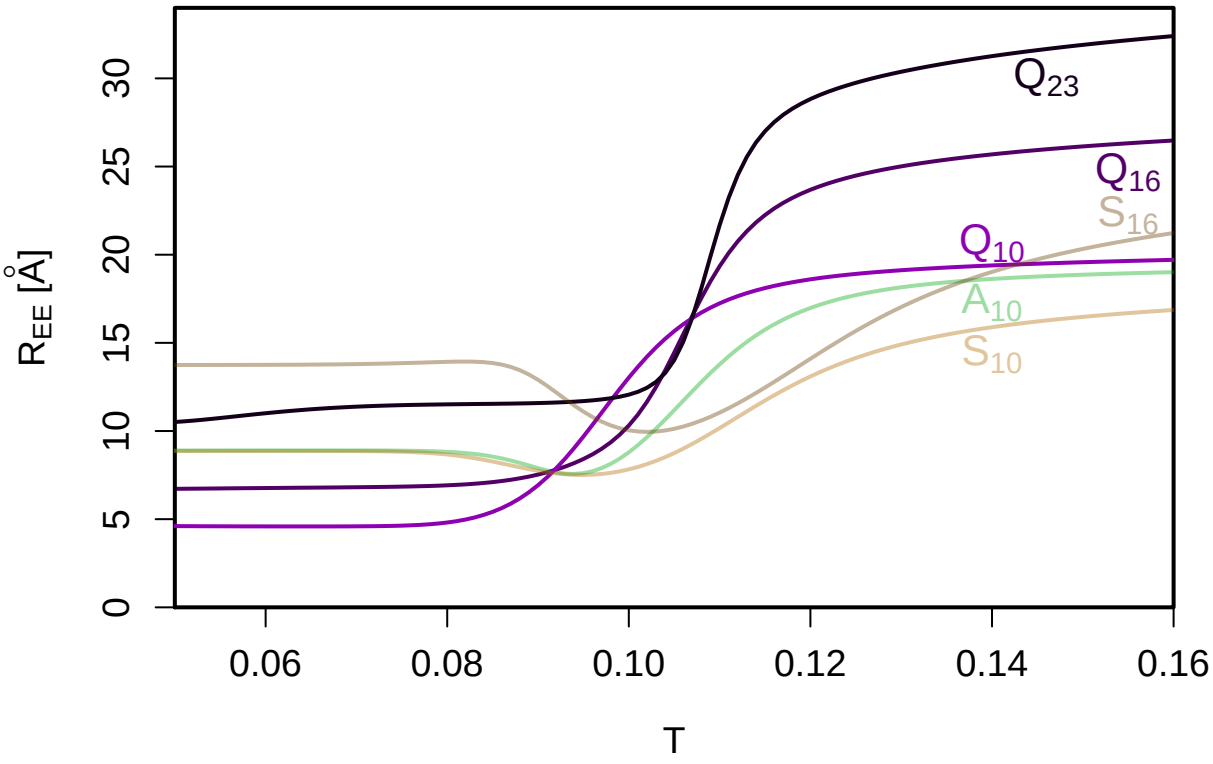

Figure 4.13 – End-to-end distance of three polyQ chains depending on temperature. polyA and polyS graphs are shown in the background for comparison.

Turning to conformational observables again, the first quantity shown in figure 4.13 is the average end-to-end distance $R_{\mathrm{EE}}(T)$, which is more informative here than the squared radius of gyration used in the polyS/polyA section. The figure shows $R_{\mathrm{EE}}(T)$ graphs of Q_{10}, Q_{16} and Q_{23} as well as A_{10}, S_{10} and S_{16} for comparison.

Similarly to $R_g^2(T)$, the end-to-end distance generally increases with temperature as the chain unfolds from a compact (folded or globular) state to an extended random coil. The distance is higher for polyQ than for polyS and polyA, indicating an increased stiffness caused by the much larger side chain, whose diameter is 3.6 Å compared to the 2.5 or 2.7 Å of serine and alanine. Otherwise, the polyQ graphs reflect the thermodynamic findings, exhibiting a single collapse from the random coil at high T to the globular/folded state at low T while polyS and polyA feature a collapse to the globule state as well as an increase of R_{EE} towards the more extended helix. Notably too, the helices of S_{10} and A_{10} have a clearly larger end-to-end distance than the low-T state of Q_{10} and even of Q_{16}, a strong hint at the formation of nonhelical configurations in polyQ chains.

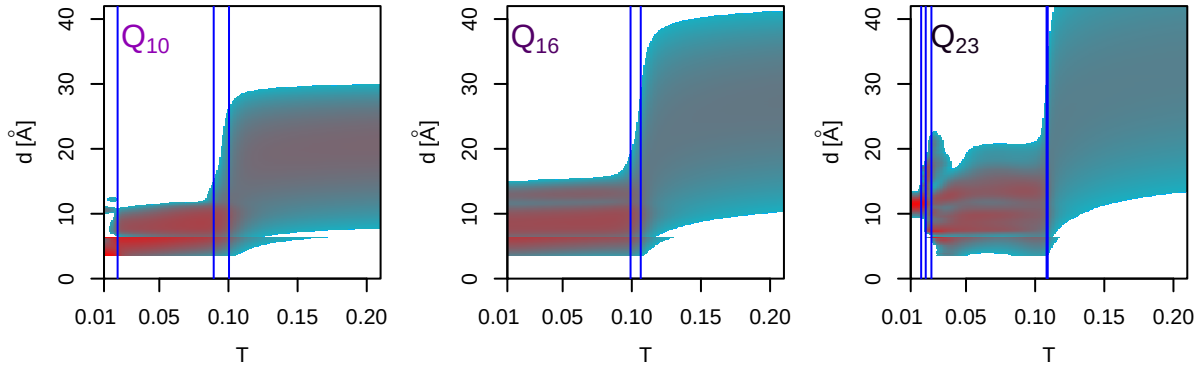

Figure 4.14 – Terminal side chain distance distributions $p(d, T)$ of three polyQ chains. Transition signatures from $c(E)$ are shown as black vertical lines.

The terminal side chain distance distributions in fig. 4.14, like those seen before in fig. 4.8, provide similar, but more detailed information. The dominant transition leads from a random coil at high T to a globule structure at lower T, which then persists for most of the temperature scale. Like in fig. 4.8, the globule is characterised by two distribution maxima above and below a fault line caused by the square well interaction. In the cases of Q_{16} and Q_{23}, a third maximum is observed around $d = 13$ Å, similar to a faint feature in the respective S_{16} and A_{16} graphs which had not been discussed there. The lower two maxima have been discussed to be an indication of the chain folding back onto itself in a kind of loop or hairpin configuration, so this third feature corresponds to a different, looser globule morphology. The maximum does not occur in the Q_{10} $p(d, T)$ graph, indicating that this kind of morphology requires a certain minimal chain length.

At very low T, both Q_{10} and Q_{23} reach a state transition which is clearly recognisable in $p(d, T)$ as well. For Q_{10}, the distribution collapses mostly towards the lowest distances, close to the hard sphere diameter of the Q side chain, but it also features a small contribution of distances around 11 Å. For Q_{23}, configurations with an end-to-end contact disappear completely at lowest T, instead a sharp maximum around 12 Å is seen here. Both of these effects are explained by certain changes in morphology, although only in the Q_{23} case the morphologies can be identified as distinct states.

In order to understand these structures, it is first necessary to identify the usual low-temperature behaviour of polyglutamine, which is shown in figures 4.15 and 4.16. Figure 4.15 contains HB matrices of Q_{10}, Q_{16} and Q_{23} at lowest T, the format being the same as in fig. 4.10. Figure 4.16 shows the corresponding snapshots of Q_{10}, Q_{16} (twice) and Q_{23}.

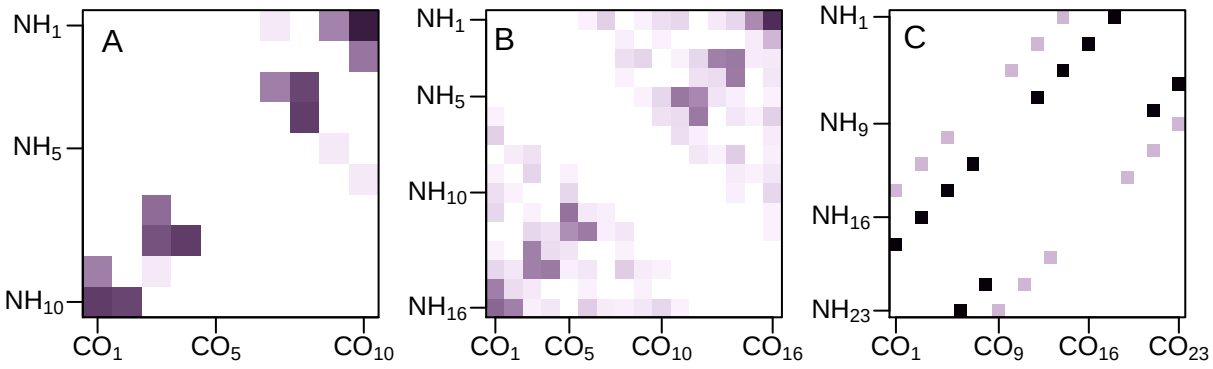

Figure 4.15 – H-Bond contact matrices of the three polyQ systems at low temperature.

Both the $N = 10$ and $N = 16$ matrices are dominated by spots on the ascending main diagonal and its side diagonals, together comprising more than 95% of configurations of Q_{10} and about 85% of Q_{16}, i.e., the sum of values in relevant neighbouring cells is 95% or 85% respectively. This summation may not be intuitive and the term "relevant" requires some explanation which follows in the next paragraphs.

The first (left) Q_{16} snapshot in fig. 4.16 shows a hairpin configuration. Such a configuration populates every other cell on the main diagonal. Like in the globule state of polyS (fig. 4.10 B), the existence of the $NH_1 \cdots CO_{16}$ H-Bond facilitates formation of the $NH_{16} \cdots CO_1$ contact, but precludes the contacts of the neighbouring residues, $NH_2 \cdots CO_{15}$ and $NH_{15} \cdots CO_2$, because the NH and CO beads of these residues are turned outwards. The next NH and CO beads (3 and 14) are turned towards each other again, allowing further H-Bonds between them. This pattern continues to the center of the chain, until contacts are either geometrically impossible or prevented by the requirement of at least three intervening residues between H-Bond partners. Thus, in this hairpin configuration, every other contact on the main diagonal is closed.

The turn in the configuration in fig. 4.16 B is relatively loose. The closest contacts are formed between residues 5 and 12 with six intervening residues. The next pair in the identified pattern would be 7–10, which cannot be formed because only two intervening residues are left here and the PRIME20 model requires at least three[3]. Hence, the highest possible number of H-Bonds in a $N = 16$ hairpin is six. The loose turn can conceivably be tightened while keeping the six H-Bonds, and indeed structures with a tighter turn are observed. The Q_{16} configuration in fig. 4.16 C shows an example of such a hairpin in which the first contact occurs between residues 3 and 16 (in contrast to the 1–16 contact seen before) and the turn lies between residues 7 and 12.

From another point of view, the structure can be regarded as a Q_{14} hairpin with two additional free residues at the N-terminus. Similarly, the additional residues can be added to the C-terminus or one to each end. Similarly, structures based on Q_{15} or Q_{13} hairpins can be formed as well without changing the number of H-Bonds. In the Q_{13} case, the innermost H-Bonds would be closed between a pair of residues i and $i + 4$, the shortest allowed spacing according to the model restriction. A Q_{12} hairpin with six H-Bonds cannot be formed, just like a Q_{16} hairpin with eight H-Bonds is impossible.

As described, the Q_{16} hairpin with a loose turn populates the main diagonal. In the Q_{15}-based hairpins, a diagonal of length 15 is populated, which can be either of the side diagonals. This population is seen in fig. 4.15 B. The Q_{14}- and Q_{13}-based hairpins either populate the more distant diagonals with corresponding length or shorter parts of the long diagonals. Thus, each hairpin lies on the main diag-

[3] If the $NH_{10} \cdots CO_7$ contact was closed, the structure would fulfil the definition of a β-turn [152, 200]. Due to the artificial restriction, such a turn can never be formed in PRIME20. Since β-turns are among the most common secondary structure elements of peptides, the usefulness of this restriction has to be reconsidered for future projects applying PRIME20. A note regarding this issue follows in the closing chapter of this thesis.

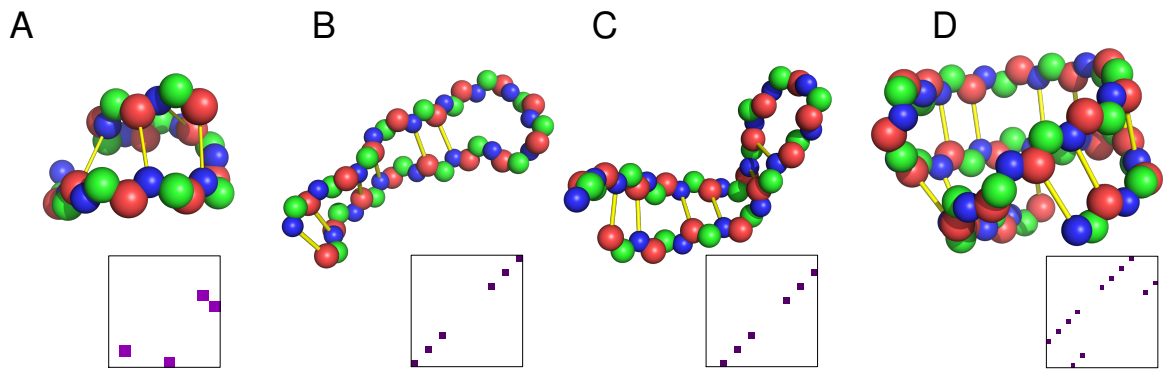

Figure 4.16 – Low-energy configurations of Q_{10} (A) Q_{16} (B and C) and Q_{23} (D) with respective H-Bond matrices. Like in earlier depictions, bead diameters are reduced to 1/4, side chains left out and H-Bonds shown as yellow sticks.

onal or one of three neighbours in either direction, which is why the aforementioned summation over neighbouring cells makes sense in order to evaluate the total hairpin content.

In the case of Q_{10}, the main diagonal (full Q_{10} hairpin) and one neighbour to each side (Q_9-based hairpins) each indicate configurations with the same number of H-Bonds, but hairpins based on Q_8 or shorter segments do not occur because they could only reach a lower number of H-Bonds. Thus, the total hairpin content relates to the sum of probabilities over a repeating "triangle" of three cells in the matrix which computes to about 95%. For Q_{16}, as discussed, the main diagonal as well as three neighbours to each side correspond to interchangeable hairpin configurations (called "relevant" neighbours in an earlier paragraph), so the triangle consists of ten cells and the hairpin content is about 85% at this temperature.

Figures 4.15 A-B and 4.16 A-C indicate that the ground state of polyQ in PRIME20 – at least up to length 16 – is a hairpin. (Q_{23} will be discussed shortly.) Another insight from the above discussion is that the number of possible H-Bonds in such a hairpin increases stepwise by two whenever 4 residues are added. In a chain of length between 9 and 12, no more than four H-Bonds are possible, from length 13 to 16, the maximum number is six, and so on. A γ-helix, identified as the ground state of polyS and polyA, contains up to $N - 4$ H-Bonds, which for all $N \geqslant 9$ is a higher number than that of a hairpin. Therefore the minimum energies reached in simulations of helical peptides are lower than those of polyQ, as seen in table 4.2.

The table lists the lowest energy observed in any simulation, U_0, and the lowest energy to which $g(U)$ converged, U_m, for all simulated peptides. For all $N \geqslant 9$, the U_m of polyQ are higher than the respective energies of polyA due to the difference between helix and hairpin, and the polyS energies tend to be even lower because polyS γ-helices are more regular than those of polyA. The U_m values of Q_9, Q_{10} and Q_{11} are almost identical because they are dominated by the maximum number of H-Bonds in their respective hairpins, which is four in all cases.

Interestingly, the lowest energies seen in simulation, U_0, do not differ much between the three polypeptides. As will be discussed along fig. 5.8 (p. 58), the γ-helical state is not entirely excluded for polyglutamines and some individual simulation runs found such configurations with correspondingly much lower potential energy. However, due to the extreme limitedness of configuration space in this state for polyQ (often called a "bottleneck" of simulation), acquiring a converged $g(U)$ for these energies would involve inappropriate effort, hence U_m is much higher. To illustrate the effort, the polyQ results shown here, despite their considerably smaller U ranges, already required more overall simulation time than the respective polyS/A results. The amount of simulation time to reach convergence in the helix state would likely have exceeded the time available in a PhD project. Furthermore, as noted before, the polyQ coil-hairpin transition temperatures translate to values between 200 and 300 K depending on solvent quality, which is roughly appropriate for biological systems. The helix transitions of polyQ on

Table 4.2 – Lowest observed potential energies U_0 and lowest potential energies U_m at which a converged $g(U)$ could be acquired of all peptides simulated with PRIME20. U_0 are values obtained from single configurations; U_m lower boundaries of the energy bins used in simulation.

	polyS		polyA		polyQ	
N	U_0	U_m	U_0	U_m	U_0	U_m
6	-3.430	-2.7				
8	-5.860	-4.0	-4.840	-4.0	-4.720	-4.5
9	-6.376	-6.0	-6.420	-5.6	-5.880	-5.2
10	-7.720	-7.0	-7.260	-6.9	-6.880	-5.3
11	-8.462	-8.3	-8.092	-7.6	-7.960	-5.3
14	-12.494	-11.8	-11.344	-10.0	-9.520	-6.5
16	-14.182	-14.0	-13.512	-12.2	-13.840	-7.4
20			-17.932	-17.5		
23					-16.880	-16.0

the other hand would lie at extremely low temperatures, which are not of any interest in experimental or real-life applications. Hence, even though the structures are somewhat interesting from the fundamental research point of view, spending a time of – in the best case – several months on them is not appropriate.

Returning to fig. 4.15, after identification and characterisation of the native polyQ hairpin, two more details call for a short description. The first is a weak signature in the Q_{10} HB matrix which does not fit into the hairpin scheme, consisting of the four interactions $NH_1 \cdots CO_7$, $NH_5 \cdots CO_9$, $NH_6 \cdots CO_{10}$ and $NH_9 \cdots CO_3$. It stems from a single simulation run reaching a non-hairpin state with equally low energy as the hairpin, which also caused the small contribution at 11 Å in the Q_{10} $p(d, T)$ graph (fig. 4.14). Unfortunately, a snapshot from this very production run does not exist, but fig. 4.16 A shows a configuration with the same energy and a somewhat similar H-Bond pattern. The configuration appears to be disordered – at least it does not contain any secondary structure elements – and its rare occurrence suggests that the hairpin is overall more variable and therefore entropically favoured at these energies. However, it becomes apparent that the hairpin is not exclusively the state of lowest energy of Q_{10}. This observation is exemplary for other short chains as well, in which disordered structures can reach energies comparable to the classical secondary structure motifs. At greater lengths ($N \gtrsim 14$), this competition does not seem to exist any more.

The other "detail" is the Q_{23} matrix, also corresponding to the low-T feature in the $p(d, T)$ graph. It consists of two distinct signatures of different configurations, easily distinguishable by colour because one occurred more frequently than the other. Both signatures are similar, which is why the description will be limited to the more prominent one. It consists of one hairpin spanning residues 1-18, clearly identified by the ascending diagonal in the HB matrix, and a second hairpin-like signature of residues 6-8 and 21-23. The structure is preferred over a single hairpin because it allows formation of 12 H-Bonds while a single Q_{23} hairpin could not exceed 10 according to the above discussion of hairpin energies. A snapshot of this configuration is found in fig. 4.16 D. It is a hairpin bending back onto itself, causing both of the chain ends to form further contacts to the residues not involved in the original hairpin in a β-sheet-like fashion. The Q_{16} chains are evidently too short to bend into such a structure, but at $N = 23$ and longer it is an option and will most likely remain as a low-temperature transition because the underlying hairpin is formed independently of this bending effect and because an extension to much larger N is easily conceivable, essentially resulting in a double-stranded kind of helix. Polyglutamines of this length have not been simulated successfully yet with PRIME20(n).

Chapter 5

PRIME20n

In the model chapter, three variants of the PRIME20 model were described, namely PRIME20 itself, its updated version PRIME20n, and a "halfway" variant created to isolate the effects of diverging parameters and named PRIME20s. The models differ by their use of so-called squeeze parameters, which describe diameter changes between beads whose contour distance along the chain is small. The discriminating features of the three models are listed in table 5.1 for reference. PRIME20 uses a size reduction to 75% between any two beads separated by three or fewer covalent bonds, PRIME20s introduces individual squeeze parameters between side chain (SC) and backbone beads separated by three, four, or (in one case) six bonds, and PRIME20n adds type-dependent backbone (BB) squeeze factors which affect beads separated by up to four bonds. These BB squeeze factors lie between 77% and 114% while the SC squeeze parameters correspond to expansions between 100% and 150%. To avoid confusion and to comply with the nomenclature used by the Hall group [39, 203], all of these modification factors will be called "squeeze factors" even if they are larger than one. It may be worth remembering that the unmodified side chain diameters are not confirmed to be identical to the ones used by the Hall group and some of the SC squeeze factors might in fact be smaller than one (cf. table 2.8).

Furthermore, the NH–CO interaction distance for H-Bond formation is larger in PRIME20n than in the other two model variants.

Table 5.1 – Differences between the three PRIME20 variants.

	PRIME20	PRIME20s	PRIME20n
Main source	Cheon, 2010 [36]	Cheon, 2015 [39]	Voegler Smith, 2001 [203]
BB squeeze factor	75%, $\leqslant$ 3 bonds	Like PRIME20	77-114%, $\leqslant$ 4 bonds
SC squeeze factor	75%, $\leqslant$ 3 bonds	Like PRIME20n	100-150%, $\leqslant$ 6 bonds
NH$\cdots$CO square well	4.2 Å	4.2 Å	4.5 Å

This chapter treats the impact of these squeeze factors on thermodynamics and structure formation of polyS, polyA and polyQ chains. The interest was sparked by the surprising observation of γ-helices in polyS and polyA simulation (cf. chapter 4). A γ-helix is a structure with regular H-Bonds between beads NH_i and CO_{i+4} (in this four-bead representation). Such a structure is shown in fig. 5.1 on the left side for a 10-residue peptide without side chains (i.e. polyglycine). The image on the right side depicts an α-helix, defined by H-Bonds between NH_i and CO_{i-4}.

Both helix types were originally described by Pauling et al. in 1951 [158] and subsequently named γ and α [156]. Two years later, Donohue [54] argued that the γ-helix might be less stable than the α-helix due to unfavourable dihedral angles and a lower number of possible van-der-Waals contacts in the core. (To illustrate this, both helices are viewed along their main axes in fig. 5.1 as well. Consistently with Donohue's argument, the pore formed by the α-helix is clearly smaller than that of the γ-helix.) Because

the number of H-Bonds is equal in both helices, the α type was expected to be the more common one in nature, and indeed the difference has turned out to be dramatic: while the α-helix is now known as the most important secondary structure element, no γ-helical structure has ever been unambiguously identified (to the author's knowledge) and its name has even been reassigned to several helix structures. It can refer to Pauling's helix, also called a 5.1_{17} helix in the notation introduced by Bragg, Kendrew and Perutz [31], but it is used for the rare 2.2_7 helix/ribbon as well [43, 128, 188], and other obscure configurations named γ-helices can be seen occasionally [145, 165, 229].

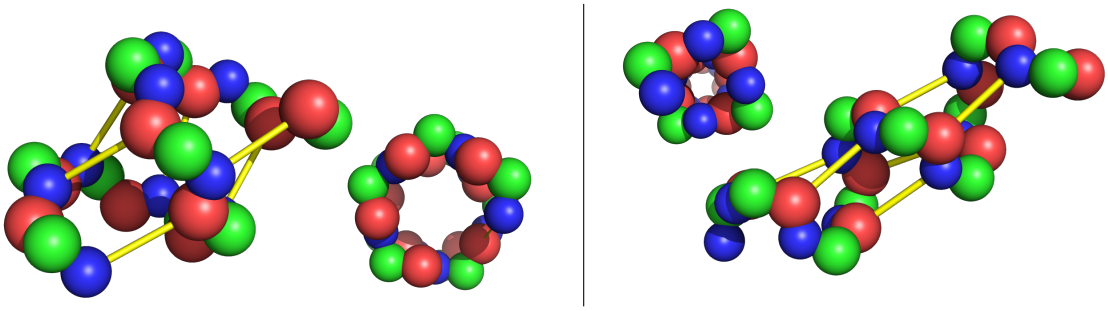

Figure 5.1 – Regular helices of type γ (left) and α (right), viewed from the side and along their main axis. The helices were constructed by assigning fixed values $(\Phi, \Psi) = (-49°, -59°)$ (α) and $(-77°, -99°)$ (γ) to all dihedral angles of a 10-residue polyglycine chain and ignoring any repulsion between beads.

This short historical excursion suggests that a protein model which reliably identifies γ-helices as native states is to be considered questionable at the very least. In PRIME20, the γ structure clearly dominates and a short investigation showed that the α-helix shown in fig. 5.1 cannot even be formed due to an unavoidable overlap between backbone beads. In a PhD thesis from 2015 [174, p. 36] Gil Rutter found the α-helical region of the Ramachandran plot to be inaccessible using PRIME, contrary to the original publications by the Hall group claiming otherwise [205] and proposed an additional size reduction to 85% for beads separated by four covalent bonds. Unbeknownst to Rutter, this additional squeeze factor is indeed close to the parameter set used in the Hall group's original PRIME simulations [203, p. 256], albeit not quite as complex. Nevertheless, the side chain squeeze parameters have not been used there.

The remainder of this chapter is structured as follows: the first section is an analysis of Ramachandran plots obtained by systematic deformation of peptides to identify how each individual squeeze parameter affects the conformational freedom. The further sections treat thermodynamics and structures of S_{10}, S_{16}, Q_{10} and Q_{16} in a comparison of the three models. Simulations of A_{10} and A_{16} have been performed too, but similar to the situation in the previous chapter, the difference to polyS is small and does not justify a lengthy explicit treatment. One exception is a curious low-T state of A_{16} to be mentioned at the very end, which has not appeared in polyS simulation.

5.1 Ramachandran plots

The aim of this section is to build up a Ramachandran plot for each of the PRIME20 variants and to understand which squeeze factors are responsible for the availability of certain secondary structures, especially the aforementioned γ- and α-helices. To this end, all Φ and Ψ dihedral angles of a polyS and a polyQ chain were varied systematically between -180° and +180° in steps of 1°. For each of the $360 \cdot 360 = 129600$ configurations obtained this way, several distances between beads were measured and formation of H-Bonds attempted.

This method provides an overview of all possible periodic, i.e. helical or zig-zag, conformations of

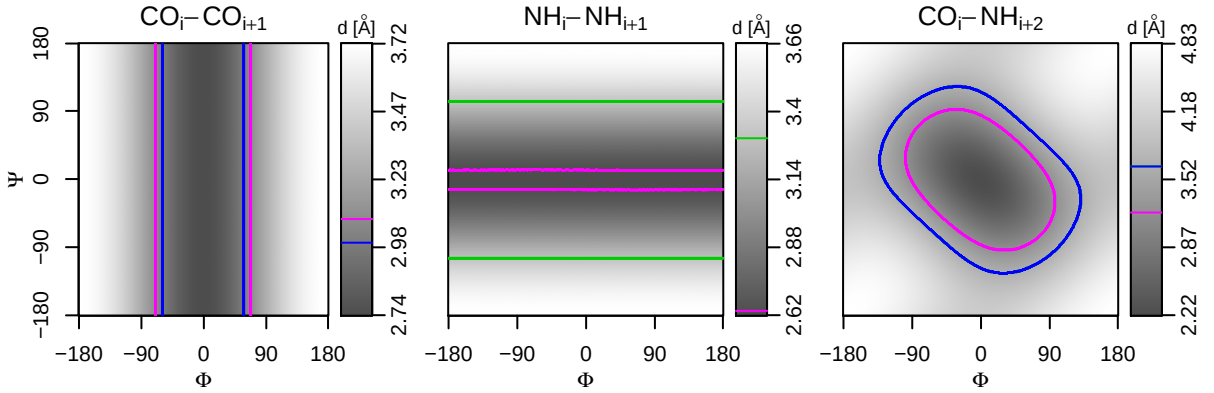

Figure 5.2 – Dependence of distances between three pairs of backbone beads on the interjacent dihedral angles Φ and Ψ. Lines indicate the respective bead sizes in PRIME20 (blue), PRIME20 ignoring the 3-bond squeeze factor (green) and PRIME20n (magenta). The area enclosed by these lines is inaccessible in the respective model.

a peptide. It should be noted that only configurations with flat peptide bonds and ideal bond angles are included here. By means of some distortion, further regions of the Ramachandran plot become accessible which are not seen in the following plots. An example of such a configuration is shown at the end of the chapter. Furthermore, non-periodic structures, for example β-turns, cannot be considered this way.

Figure 5.2 shows the dependence of three backbone bead distances on the dihedral angles, namely the distance between two neighbouring CO beads (left), two neighbouring NH beads (middle) and between a CO bead and the NH bead two residues further along the chain. Light colours stand for long distances according to the colour scales beside each plot. The coloured lines mark the respective bead diameters in PRIME20[1] (blue), in PRIME20n (magenta), and in a hypothetical model using only the original bead sizes without any squeeze factors (green). The regions enclosed by these lines would be unavailable to molecules in the respective models.

The CO$_i$–CO$_{i+1}$ distance is independent of Ψ because the dihedral angle corresponds to a rotation about the C$_\alpha$–CO bond and leaves the relative positions of the two CO beads unaffected. Therefore the whole plot has a visually vertical layout, including vertical cut-off lines. The lines of PRIME20 and PRIME20n are fairly close to each other, corresponding to distances of 3 Å and 3.0852 Å (sic) respectively. The original bead size is 4 Å, but the CO$_i$–CO$_{i+1}$ distance never exceeds 3.72 Å, hence the green line is not drawn. Without any squeeze factors, all configurations would be illegal due to this overlap.

The second image shows the NH$_i$–NH$_{i+1}$ distance, which is independent of Φ in the same way in which the CO$_i$–CO$_{i+1}$ plot is independent of Ψ. The green line signifying the original bead size of 3.3 Å is visible this time, but the blue line for the PRIME20 cut-off disappears. The cut-off lies at 2.475 Å, but the lowest distance here is 2.62 Å. This means that an NH$_i$–NH$_{i+1}$ overlap can never occur in PRIME20 (barring shortened bond lengths). In PRIME20n however, the cut-off lies at 2.64 Å, making a thin band of Φ angles between -15° and +11° illegal.

The CO$_i$–NH$_{i+2}$ distance in the rightmost plot depends on both interlying dihedral angles, Φ_{i+1} and Ψ_{i+1}. Because the beads are separated by four bonds, PRIME20 uses the original diameters, hence the blue and green lines coincide. In the first two plots, PRIME20n was found to be slightly more restrictive than PRIME20, but here the relation is the inverse: due to the lack of a 4-bond squeeze factor in PRIME20, the PRIME20n cut-off at about 3.07 Å leaves more flexibility than the PRIME20 cut-off at 3.65 Å. This will prove to be a crucial difference between the models regarding α-helix formation.

[1]The backbone geometries of PRIME20 and PRIME20s are identical. For the sake of readability, PRIME20s will not be named explicitly in this context.

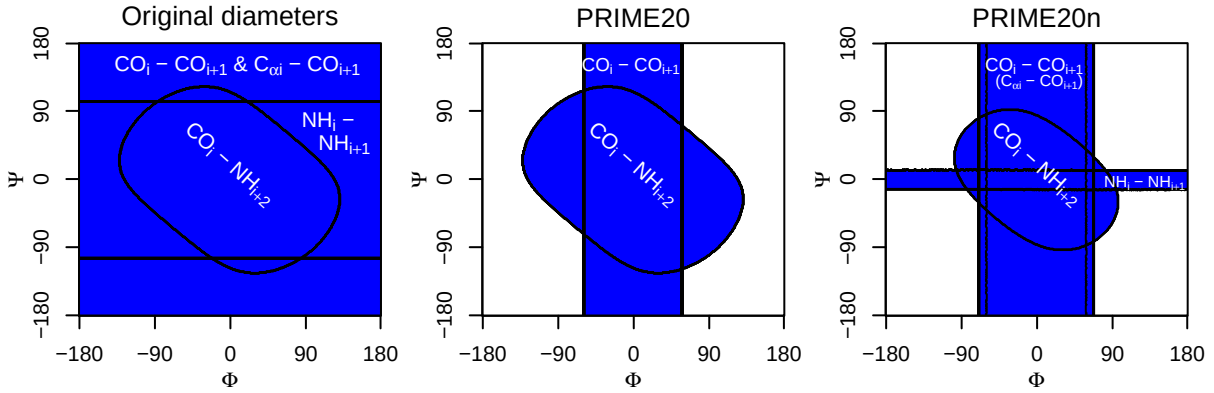

Figure 5.3 – Effect of backbone bead repulsions on the Ramachandran map in PRIME20 without the 3-bond squeeze factor, PRIME20, and PRIME20n. The blue area is inaccessible; labels indicate which individual repulsion event restricts which part of the map.

A combination of the three pictures in figure 5.2 produces the backbone Ramachandran plots shown in fig. 5.3 for the original (unsqueezed) diameters, PRIME20, and PRIME20n. Black lines mark the borders between regions which are legal or illegal due to certain distances identified by text labels, corresponding to the coloured lines in fig. 5.2. The legal region is coloured white, the illegal region blue. In addition to the three distances seen in fig. 5.2, two further squeeze factors affecting $C_{\alpha i}$–CO_{i+1} and NH_i–$C_{\alpha i+1}$ were considered too (see table 2.6). However, the first does not restrict conformation space any more than CO_i–CO_{i+1} and the second never has an effect at all.

The cut-off distances and lines have been discussed in the above paragraphs already: using the original bead sizes, all configurations are illegal due to the CO_i–CO_{i+1} and $C_{\alpha i}$–CO_{i+1} overlaps. In PRIME20 and PRIME20n, these overlaps cause a vertical illegal band in the plot which is overlaid by the CO_i–NH_{i+2} oval. In PRIME20n, the thin horizontal band caused by the NH_i–NH_{i+1} overlap is seen which does not occur in PRIME20.

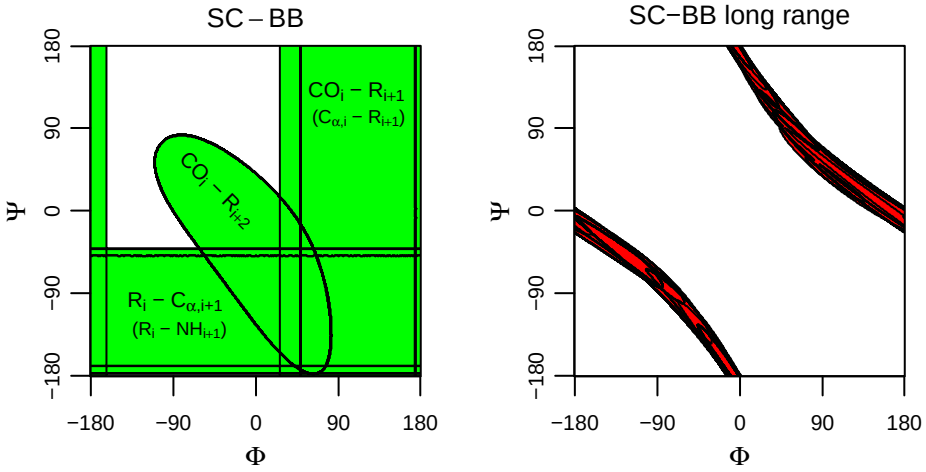

Figure 5.4 – Effect of side chain bead repulsions on the Ramachandran map in PRIME20n for polyserine. Interactions of side chains with close neighbours along the backbone are shown in the left map, longer-distance interactions on the right. The inaccessible areas are coloured green or red.

In addition to the backbone overlaps, the side chain beads restrict conformation space as well. Two unrelated effects caused by the side chains are shown in fig. 5.4. The first, in the left image, relates to the squeeze factors *sqz*6-10 (table 2.7) which govern the three- and four-bond distances between side chain (SC) and backbone (BB) beads as well as the six-bond distance CO_i–R_{i+2}; the second effect, on the right,

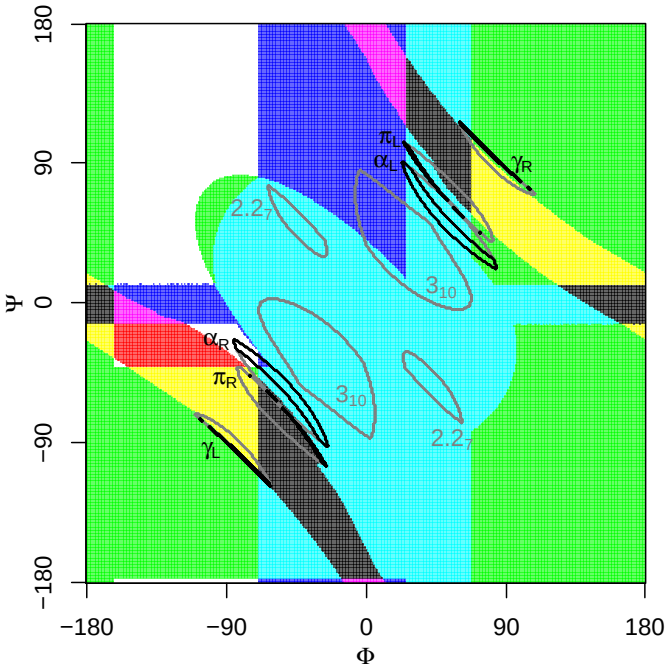

Figure 5.5 – Complete Ramachandran plot for a polyserine chain in PRIME20n. The white area is legal; colour indicates which type of repulsion forbids the respective (Φ, Ψ) set: red for side chain interactions, green for interactions between side chain and backbone, blue for backbone interactions. Overlapping regions are coloured corresponding to additive mixing, i.e. yellow=red+green, turquoise=blue+green, magenta=red+blue, black=red+blue+green. Black outlines show the regions in which the defining H-Bonds of different helix types would be closed (assuming the structure itself was legal). Regions in which the H-Bonds are hindered by auxiliary interactions are outlined in grey.

describes several overlaps between beads separated by a greater number of bonds. Like in fig. 5.3, black lines and text labels indicate overlap boundaries. The legal regions are coloured white and the illegal regions green or red. Both images refer to a polyserine chain. The qualitative shapes are the same for polyglutamine, but their positions and sizes change, as will be seen later.

The left-hand figure pertains to PRIME20n/s. The same figure for PRIME20 would be an empty white field because the hard-sphere diameters in PRIME20 are much smaller than the squeeze parameters in PRIME20s/n and never experience one of the considered overlaps. For PRIME20n/s however, the side chain squeeze parameters play a huge role regarding conformational freedom. A broad vertical illegal band is caused by the overlap of the side chain with the preceding CO bead (and another overlap with the C_α bead), and a horizontal band by an overlap with the next C_α (and NH) bead. The six-bond effect depends on two dihedral angles and produces an asymmetric oval shape in the center of the plot, similar to the CO_i–NH_{i+2} overlap. Only 23% of the plot are white, i.e. legal.

The second plot in fig. 5.4, treating "long-range" interactions between a side chain bead and backbone beads separated by up to five residues along the chain, is identical in all model variants because these interactions are not governed by squeeze parameters. Unlike the previous figures, in which dimer effects depending only on a single (Φ, Ψ) pair were treated, the idea here is that a side chain in a helical configuration may collide with a backbone bead in the previous or following helix turn. Such a collision depends on multiple dihedral angles, all assumed to be equal. The impact of these collisions is limited to a thin band compared to the previous images, but as the following figures will show, the shape of this band, depending on the side chain type, is crucial.

A combination of the partial effects shown in figures 5.3 and 5.4 leads to a complete Ramachandran plot of legal, regular structures. Fig. 5.5 shows this Ramachandran plot for a PRIME20n polyserine

chain. It consists of many coloured areas which are derived from the previous plots: blue signifies regions which are illegal due to an overlap of backbone beads (from fig. 5.3), green regions made illegal by interactions between side chain and neighbouring backbone beads (left side of fig. 5.4). These regions are essentially inaccessible for all types of configurations, helical and nonhelical. Red marks the bands caused by longer-distance interactions between side chain and backbone in a regular structure. These angles are available to single amino acids or dimers, but they cannot be repeated to form a complete helix turn. Further colours in the figure stem from additive mixing: if at least one backbone interaction (blue) and one side chain interaction (green) make a (Φ, Ψ) pair inaccessible, the respective region of the matrix is coloured turquoise. Similarly, magenta is the combination of long-range side chain repulsions and backbone interactions (red+blue) and yellow the overlap of short-range and long-range side chain to backbone repulsions (green+red). Finally, the black area corresponds to angles at which all three kinds of repulsions are active, and white marks the legal region.

In addition to the coloured areas, helical regions are drawn in the diagram as black and grey outlines. Black surrounds the areas in which certain types of H-Bonds are formed according to the model rules (using a square well diameter of 4.5 Å). The areas are labeled α for $NH_i \cdots CO_{i-4}$ H-Bonds, γ for $NH_i \cdots CO_{i+4}$ and π for the $NH_i \cdots CO_{i-5}$ type, with an index R or L indicating right- or left-handedness of the helix. Grey outlines surround the areas in which these H-Bonds could be formed, were they not hindered by auxiliary interactions[2]. Two helix types only appear in this way, the 3_{10} helix with an $NH_i \cdots CO_{i-3}$ H-Bond pattern, which is observed occasionally in nature, and the exceedingly rare 2.2_7 helix with $NH_i \cdots CO_{i-2}$ H-Bonds. As a side note, the $NH_i \cdots CO_{i+3}$ H-Bond scheme, which would correspond to a 3.4_{14} helix, does not occur at all. Like the γ-helix, it has been identified as an unfavourable configuration by Donohue [54] and never observed since then. Further imaginable helices with $NH_i \cdots CO_{i+5}$ or $NH_i \cdots CO_{i+2}$ patterns do not occur either and they are not even mentioned in Donohue's paper.

The legal region in fig. 5.5 is mostly rectangular, ranging from $(\Phi, \Psi) = (-163°, 180°)$ in the top left corner to $(67°, -42°)$ in the bottom right. The rectangle is further deformed by the horizontal NH_i–NH_{i+1} repulsion band, by the oval overlap shape caused by the CO_i–R_{i+2} and CO_i–NH_{i+2} repulsions and by the long-range side chain repulsion line overlaying most of the $\Psi < 0$ section. Notably however, a significant part of the α_R-helix region remains legal. Furthermore, the right-handed π and 3_{10} regions are not far from the borders and it is conceivable that these configurations may occur

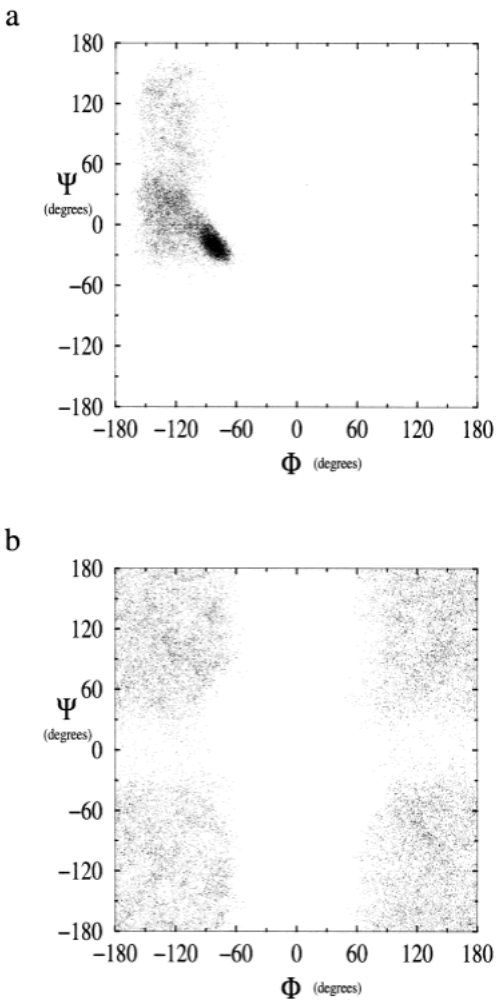

Figure 5.6 – Ramachandran plots of (a) a peptide with side chains and (b) a polyglycine chain without side chains in the PRIME model (Voegler Smith et al., 2001 [205, fig. 5]).

[2] As described in the model section, these H-Bonds appear to be formed in DMD simulation, but do not yield an energy gain and are easily broken again. They cannot be closed at all in the MC interpretation.

by virtue of some deformation. The γ_R helix as well as all left-handed types are virtually impossible to build according to this image. This makes the Ramachandran plot a quite convincing one – the common helix structures are available (α_R, single turns of π_R, $3_{10,R}$) and those which are not usually observed are illegal in this plot. Non-repeating structures like β-sheets or random coils are typically found in the top left region, which is legal as well.

As a further validation, the available section of the plot and the location of the α-helix region correspond quite well to depictions found in literature [65, 80] and most importantly the Ramachandran plot published by Voegler Smith et al. [205] (fig. 5.6 a). The side chain there is not a serine side chain – in fact, it is not related to any PRIME20n side chain because the PRIME parameters differ – but the accessible region is qualitatively the same. The α_R-helical region is strongly populated in this plot due to energetic preferences, and it is larger than the region labelled α_R in fig. 5.5 due to the allowed deformation of bond lengths and angles. For the same reason, the horizontal NH_i–NH_{i+1} band is not visible at all.

Figure 5.6 b shows simulation results from a polyglycine chain, i.e. a peptide without side chains. This figure corresponds to the PRIME20n backbone Ramachandran plot in fig. 5.3, and indeed the features are similar, perhaps the most notable difference being the somewhat broader NH_i–NH_{i+1} band in Voegler Smith's Ramachandran plot compared to fig. 5.3. However, the band is not entirely unpopulated, once again highlighting the flexible behaviour of the bond lengths.

Now, in order to understand the impact of squeeze factors, the Ramachandran plot of a PRIME20n polyS chain (fig. 5.5) can be compared to those of the other model variants, shown in fig. 5.7. The figure contains, from top to bottom, the polyS Ramachandran plots in PRIME20s, PRIME20 and a fourth variant, to be called $PRIME20_4$ in the following paragraphs[3]. With two sets of squeeze parameters, both of which can be applied independently, a total of four models can be created: PRIME20n uses both sets, PRIME20 neither, PRIME20s only the side chain squeeze parameters and $PRIME20_4$ only those of the backbone.

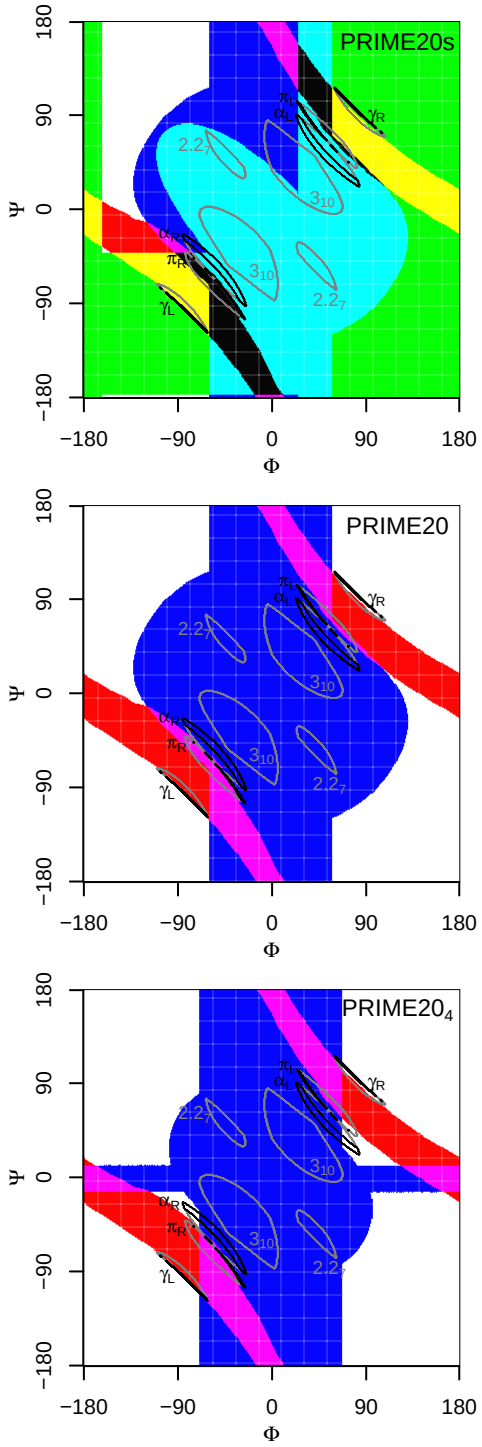

Figure 5.7 – Ramachandran plots of polyserine in PRIME20s, PRIME20 and another model variant, $PRIME20_4$.

In the PRIME20s plot, compared to PRIME20n, the increased range of the CO_i–NH_{i+2} overlap is the most important difference. As seen in fig. 5.3 already, it inflates the central ellipsoid shape. This shape reduces the overall available area, but most importantly it

[3]The $PRIME20_4$ model is included here for the sake of completeness, but it is just as hypothetical as PRIME20s and has not been used in simulation.

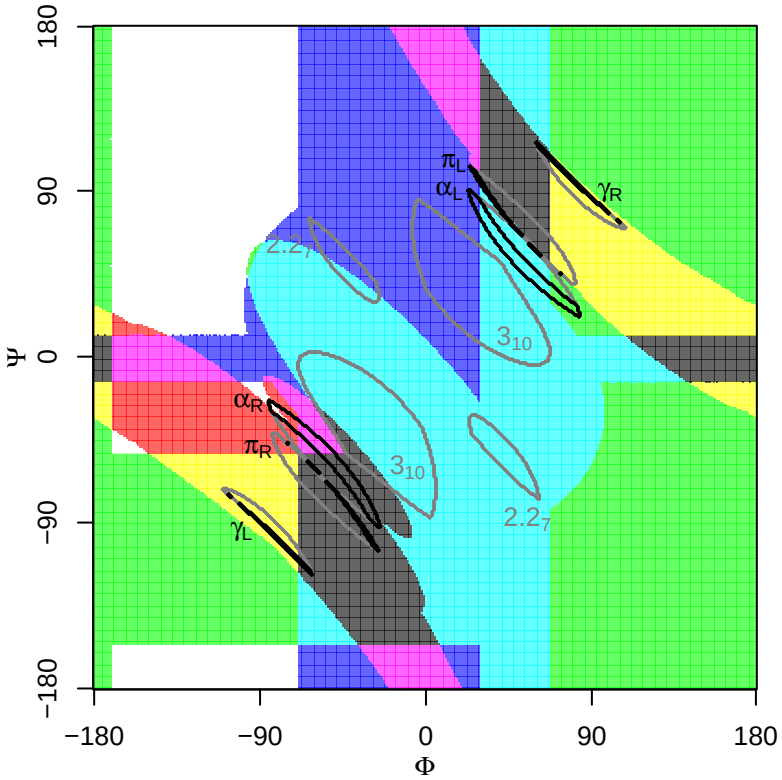

Figure 5.8 – Complete Ramachandran plot of a PRIME20n polyglutamine chain.

overlays the α_R-helix region, making this configuration inaccessible to a PRIME20s chain. The NH_i–NH_{i+1} and CO_i–CO_{i+1} interactions cover a reduced area because the squeeze factor of 75% is smaller than the respective factors of 80% and 77.13% used in PRIME20n. This means that the horizontal block (caused by the NH_i–NH_{i+1} overlap) disappears completely and the vertical block (CO_i–CO_{i+1}) is slightly smaller, but they do not affect the overall layout of the plot as much as the CO_i–NH_{i+2} oval.

Removing the side chain squeeze parameters from PRIME20s leads to PRIME20 and the middle plot in fig. 5.7. Because almost all side chain squeeze factors are larger than 1, their removal greatly increases flexibility. The green area representing the squeezed side chain overlaps disappears completely. Because the backbone is the same as in PRIME20s, the α-helix regions are still inaccessible, but unlike in PRIME20s the region assigned to γ-helices is legal here. It lies close to the boundaries of the red band signifying long-range side chain interactions, so these interactions may still restrict the formation of such helices, but according to the plot – and to the observations in the previous chapter – such helices are available to some extent. Interestingly, while in PRIME20n (as in nature) only right-handed α-helices can be formed, both the γ_R and γ_L regions are accessible in PRIME20, essentially allowing the formation of two different structures with equal energy.

Finally, $PRIME20_4$ means the "return" of the backbone squeeze factors. The central ellipse is smaller than in PRIME20 and frees up the α_R-helix region again. Because the side chain squeeze parameters are not used here, the γ regions remain legal. Even more, the α_L region is available in this model variant as well because the asymmetry in PRIME20n is only caused by the side chains, not by the backbone. The thermodynamic behaviour of this model could be expected to be rather complex, with a competition between several helix types at low energy, which are only – if at all – distinguished by a slight difference in side chain attractions. Furthermore, π-helical structures appear to be available as well. With only one less H-Bond compared to α- and γ-helices, their energy would not be much higher and they would probably play a role in simulation. However, such simulations did not take place because $PRIME20_4$

is just as hypothetical as PRIME20s and the much more restricted conformation space in PRIME20s promised more insight into the effects caused by the squeeze parameters.

If the serine side chain is replaced by a different one, the Ramachandran plot will change accordingly. The blue area of backbone interactions will be identical, but the green and red areas describing side chain effects will change their size or position. The Ramachandran plot of polyalanine is found in the appendix (fig. B.1) for reference. Despite the different side chain sizes and positions making polyA less flexible than polyS, the relevant features, especially the availability of the α-helix, are the same. In the case of polyglutamine, seen in fig. 5.8, the differences are more significant.

Because the glutamine side chain is longer than that of serine, its center of mass and thus the side chain bead in PRIME20(n) lies further away from the backbone. This diminishes the impact of the interactions between the i-th side chain bead and the $(i+1)$-th NH and C_α beads, narrowing the horizontal band to $\Psi = [-156°, -53°]$ compared to $\Psi = [-177°, -42°]$ of serine, extending the β region (see fig. 5.2 for reference) and would free up some space for a π_R helix, were this helix not subjected to the long-distance side chain interactions too.

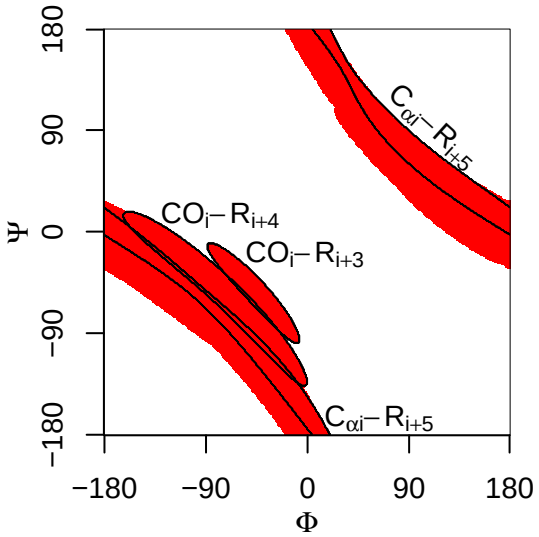

Figure 5.9 – Contribution of long-distance interactions between side chains and backbone to the polyQ Ramachandran plot (fig. 5.8). Three of these interactions are marked by black outlines.

These are changed due to the considerably larger side chain bead of glutamine with a diameter of 3.6 Å compared to 2.5 Å of serine, broadening the red-coloured band of such interactions. An elliptical shape caused by the CO_i–R_{i+3} interaction is separated from the rest of the band (cf. fig. 5.9) and impedes α-helix formation. Only a small spot in the α_R-helix region stays free. Furthermore, the π_R-helix, which would be made available by the reduced horizontal band as noted above, is hindered by the increased CO_i–R_{i+4} overlap. Finally, both γ-helix regions are inaccessible mostly due to the repulsion between $C_{\alpha i}$ and R_{i+5}.

To conclude the section, evidently the most relevant squeezed interactions are the CO_i–NH_{i+2} overlap and some of the side chain parameters. The CO_i–NH_{i+2} repulsion hinders the α-helix in PRIME20 while the side chain overlaps render the γ-helix illegal in PRIME20n. This leads to the observation of polyS and polyA γ-helices in PRIME20, and both can be expected to form α-helices in PRIME20n.

According to the Ramachandran plots, polyQ is incapable of forming γ-helices and very unlikely to form α-helices in the respective models. In PRIME20, this leads to hairpin low-temperature states, and the configurations in PRIME20n should look similar. The situation in the PRIME20s model is practically the same, with all helix types being illegal regardless of the individual amino acid type, so all peptides are expected to form hairpin structures here.

As discussed along table 4.2 (p. 50), polyQ γ-helices are actually observed – albeit rarely – by virtue of bond length fluctuations. This underlines that these Ramachandran plots do not set fixed rules, but only indicate which structures can be expected for which peptide. Because of the large number of individual bond lengths, all of which could affect the plots in some way, a comprehensive discussion of their influence is not possible.

5.2 Simulation results: Thermodynamics

The first section of this chapter served to identify which low-temperature structures can be expected in simulation of the three models PRIME20, PRIME20s and PRIME20n. The simulation results in the following sections will verify these expectations, leading to an overall consistent picture of the model variants. Some configurational details will be shown which the Ramachandran plots could not predict. In all three variants, simulations have been performed for polyS, polyQ and polyA chains of lengths 10 and 16. Because the polyA results do not differ much from those of polyS, the discussion will mostly be limited to the four systems S_{10}, S_{16}, Q_{10} and Q_{16}.

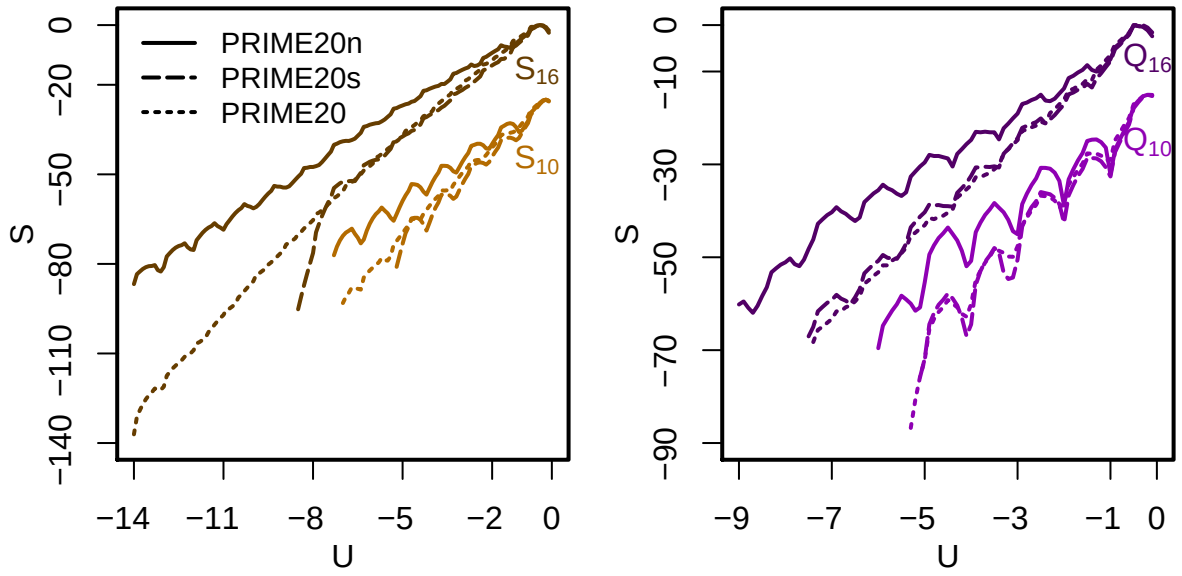

Figure 5.10 – Configurational entropies of polyS and polyQ in PRIME20, PRIME20s and PRIME20n. Systems are distinguished by colour, models by line type.

Like in the previous chapter, analysis begins at the configurational entropy $S(U)$, the primary result of an SAMC simulation. Fig. 5.10 depicts these entropies for each of the four peptides in three models. Short-dashed lines signify PRIME20, long-dashed lines PRIME20s and continuous lines PRIME20n. The $N = 10$ graphs are shifted downwards to avoid intersections.

An obvious feature distinguishing the graphs is the visually perceived slope, which is smaller in PRIME20n than in the other two systems. PRIME20 and PRIME20s are similar in this respect. The visual observation is somewhat quantifiable by comparison of codomains: over identical energy ranges, the PRIME20n and PRIME20 graphs of S_{16} span entropy intervals of $[-87, 0]$ and $[-137, 0]$ respectively and the other systems behave similarly. Because the derivative of entropy is the inverse temperature (eq. (3.22)), this smaller slope translates to a main $C(T)$ peak at a much higher temperature in PRIME20n than in the other models, which will be seen in fig. 5.11.

The configurational entropy quantifies the portion of configuration space which corresponds to a given potential energy. Because the available configurations, especially at low energies, are more restricted in PRIME20s than in PRIME20n, it is intuitive that the entropy in PRIME20s at low U is lower and therefore the graph must be steeper.

PRIME20 on the other hand has a much larger legal configuration space according to the Ramachandran plots (fig. 5.8), but the $S(U)$ slope is still larger than in PRIME20n. To explain this difference, table 5.2 lists the accessible Ramachandran plot area A in each model variant as well as the α- and γ-helical fractions of this area, A_α/A and A_γ/A. A PRIME20 chain has access to about 45% of all possible

Table 5.2 – Legal Ramachandran plot area A and α- and γ-helical fractions $A_{\alpha,\gamma}/A$ of polyS, polyA and polyQ in the four PRIME20 variants (including PRIME20$_4$, as seen in fig. 5.7). Only columns with nonzero entries are shown.

	PRIME20		PRIME20s	PRIME20n		PRIME20$_4$		
	A	A_γ/A	A	A	A_α/A	A	A_α/A	A_γ/A
polyS	45%	0.20%	10%	11%	0.70%	48%	0.31%	0.16%
polyA	46%	0.22%	5%	7%	0.43%	49%	0.31%	0.16%
polyQ	42%	0	12%	14%	0.13%	46%	0.20%	0

(Φ, Ψ) pairs while a chain in PRIME20n can use only about 10% of the plot area. Of those 45%, about 0.2% belong to the γ-helical regions of PRIME20. In PRIME20n, the fraction of α-helical (Φ, Ψ) pairs is considerably larger, making up 0.7% for polyS and 0.43% for polyA. Hence, conformational freedom in the α-helical state of PRIME20n is higher than in the γ-helical state of PRIME20, when compared to the respective random coil states (which are assumed to cover the entire accessible plot areas).

This way, arguments could be found why the $S(U)$ graphs of both PRIME20 and PRIME20s are steeper than those of PRIME20n; however, the arguments are not directly related to each other and the similar slopes of the PRIME20/PRIME20s graphs must be regarded as a coincidence.

Table 5.3 – Lowest observed potential energies, U_0, and lowest potential energies at which a converged $g(U)$ could be acquired, U_m, of the peptides simulated with all three PRIME20 variants.

	PRIME20		PRIME20s		PRIME20n	
	U_0	U_m	U_0	U_m	U_0	U_m
S_{10}	-7.720	-7.0	-5.548	-5.2	-7.204	-7.3
A_{10}	-7.260	-6.9	-5.840	-4.9	-6.756	-6.6
Q_{10}	-6.880	-5.3	-5.200	-5.1	-7.120	-6.0
S_{16}	-14.182	-14.0	-10.268	-8.5	-14.236	-14.0
A_{16}	-13.512	-12.2	-9.596	-8.6	-13.176	-13.0
Q_{16}	-13.840	-7.4	-9.160	-7.5	-10.480	-9.0

Another difference between the $S(U)$ graphs in fig. 5.10 pertains to the range of available energies. The PRIME20s graphs of polyS span significantly shorter U ranges than the comparable PRIME20 and PRIME20n graphs. The difference is smaller for polyQ; here only the energy range of Q_{16} in PRIME20n is longer than the other ones. To quantify and complete this observation, the lowest reached energies (U_0) and the lowest energies at which a converged density of states could be obtained (U_m) are listed in table 5.3. As the table shows, the PRIME20s energies are generally higher than those in PRIME20 and PRIME20n. The latter two are not clearly distinguished in this regard; their largest difference is found in Q_{16}, where PRIME20n converged better than PRIME20 in the same simulation time (as noted before), but a single PRIME20 run reached a γ-helical state with correspondingly low energy. The lowest energy observed across all other Q_{16} PRIME20 runs is -10.880, close to the PRIME20n value.

Following the discussion of Ramachandran plots, chains in PRIME20 and PRIME20n are expected to form γ- or α-helices as their ground states, both of which contain up to $N-4$ H-Bonds. This expectation is corroborated by the energies of polyA and polyS in these models because all of these energies lie slightly below $-(N-4)$. PRIME20s chains as well as polyQ in all models are expected to form hairpins instead. The lowest converged energies U_m lie around $-N/2$ in all of these cases, again supporting the expectations. The U_0 of polyQ tend to be lower due to helical states, which are reached on occasion, but

not consistently across all simulations, thus they do not suffice to produce a converged density of states. The only observation that cannot be explained (at a hypothetical level) by the Ramachandran plots is the lower U_m of Q_{16} in PRIME20n compared to PRIME20(s). It belongs to a specific configuration, which is non-repeating and therefore not caught by the Ramachandran plots, but will be identified in the discussion of fig. 5.20.

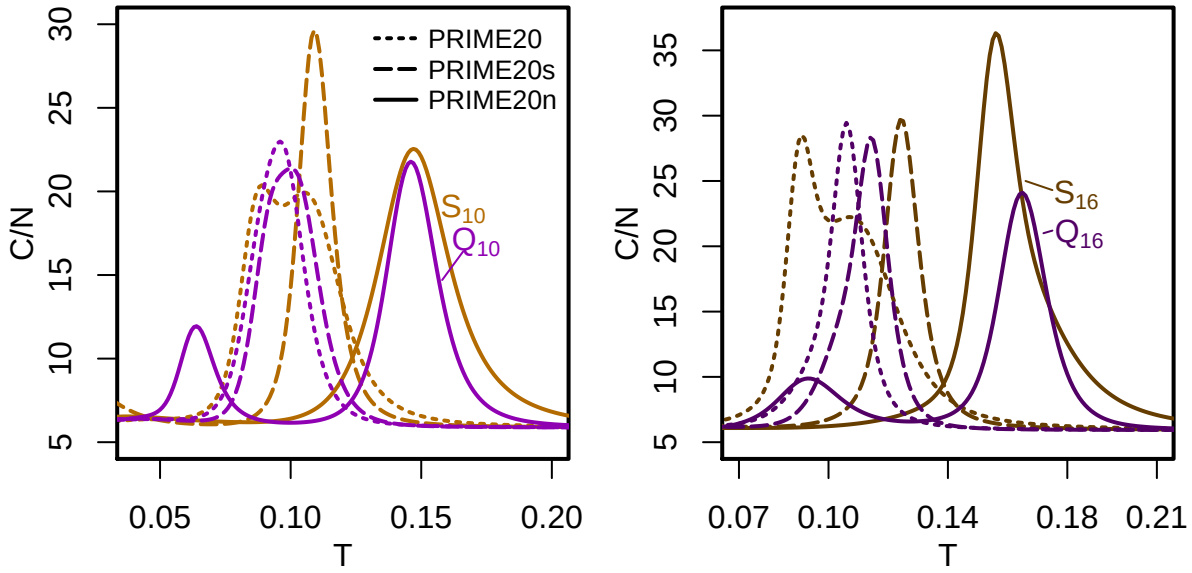

Figure 5.11 – Canonical heat capacities of polyS and polyQ in PRIME20, PRIME20s and PRIME20n.

As already noted, the heat capacities, shown in fig. 5.11, experience a shift towards higher temperatures in PRIME20n relative to PRIME20(s). This observation holds for both $N = 10$ (left) and $N = 16$ (right), and for both polyS (brown lines) and polyQ (purple). Apart from this shift, the overall shapes of the graphs are similar, containing one dominant maximum which corresponds to a collapse or folding transition, and sometimes a smaller maximum which indicates a two-step folding process from the random coil to the ground state.

Both of the PRIME20 polyserine graphs exhibit a double peak structure caused by two transitions between random coil, disordered globule and γ-helix, as already seen in the previous chapter (fig. 4.5). The polyglutamine graphs only feature a single peak for the coil-hairpin transition.

In PRIME20s, all $C(T)$ curves contain a single maximum. In absence of a helix state, this maximum is expected to describe a transition between the random coil and a hairpin structure. However, it should be noted that the low-T regions have been cut off to enhance clarity and further effects at $T < 0.05$ will be described later.

In PRIME20n, it is polyQ which features a second $C(T)$ peak within the plot window, again indicating the existence of another folded state. Such a second maximum does not exist in the polyS graphs. At first glance, polyS appears to follow a two-state folding process between random coil and α-helix, different from the three-state process in PRIME20, but an asymmetry in the $C(T)$ graphs reveals the existence of a collapse transition whose signature is merged into the folding peak. Its existence can be confirmed by microcanonical analysis, where two maxima in $c(E)$ are found. (The $c(E)$ graphs are shown in the appendix in fig. B.2.) The corresponding temperatures are $T = \{0.139, 0.150\}$ (S_{10}) or $T = \{0.156, 0.170\}$ (S_{16}), all of which lie well within the respective $C(T)$ peak widths and are therefore not resolved in the canonical ensemble.

5.3 Simulation results: H-Bonds

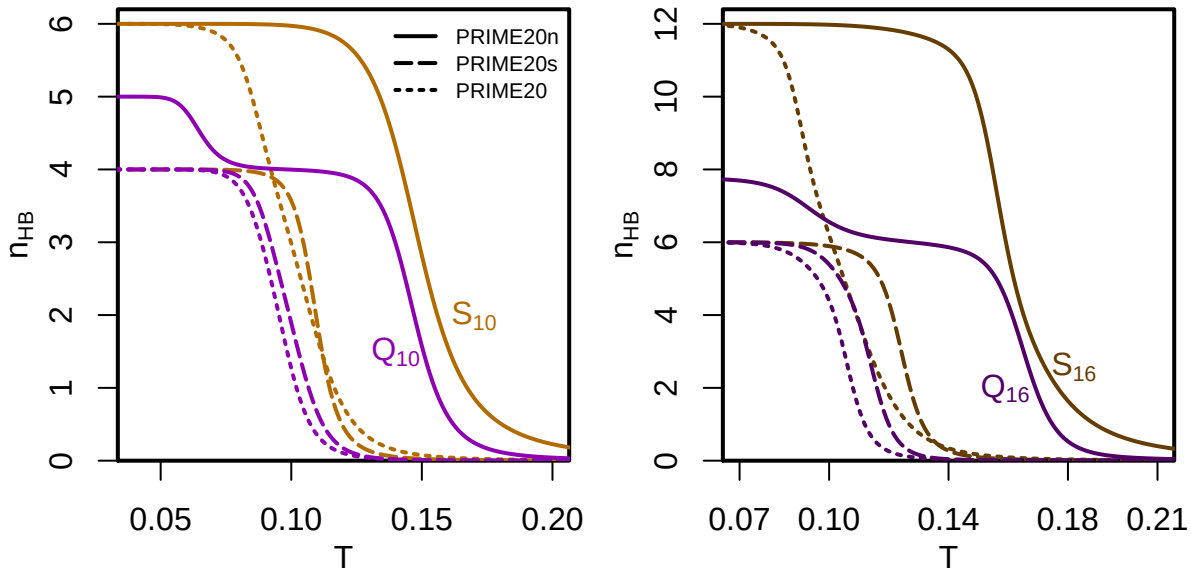

Figure 5.12 – Average number of H-Bonds of polyS and polyQ versus temperature in PRIME20, PRIME20s and PRIME20n.

The thermodynamic behaviour of the investigated peptides is not very complex, in many cases consisting of a single notable transition signature. These transitions are conjectured to distinguish certain hairpin, helix or coil states based on results from the preceding chapter and from the Ramachandran plots, but these states are still to be identified conclusively.

The most powerful tool of structure identification used in the previous chapter are the H-Bond contact matrices. They provide a clear picture of the configurations at fixed temperature (or energy), but they are not well suited for a T-dependent discussion. To this end, lower-dimensional quantities like the total number of H-Bonds $n_{\mathrm{HB}}(T)$ are more useful.

The number of H-Bonds of polyS and polyQ in all three model variants is shown as a function of temperature in figure 5.12. The behaviour is largely the same in both subfigures. Three graphs – namely those of polyQ in PRIME20 and PRIME20s as well as of polyS in PRIME20s – consist of a single jump from $n_{\mathrm{HB}} = 0$ at high temperatures to $n_{\mathrm{HB}} = 4$ (for $N = 10$) or 6 (16) at low T. The position of this jump varies between the graphs, but it equals the maximum of $C(T)$ for all of them. The PRIME20 and PRIME20n versions of polyS feature similar jumps, ending at 6 or 12 rather than 4 or 6. Finally, the polyQ graphs in PRIME20n first increase from 0 H-Bonds to a plateau at 4 or 6 and then – at the temperature of the second heat capacity peak – further to 5 or 8.

If the aforementioned assumptions about formed structures hold, the large increase of the polyS graphs should be caused by helix formation. The smaller increases seen in the polyQ and PRIME20s graphs can supposedly be attributed to a collapse into globule or hairpin states, and the reason for the second transition of polyQ in PRIME20n is still unclear. The lower number of H-Bonds compared to the helices suggests a nonhelical conformation, however the existence of an α-helical fragment has not been ruled out by the Ramachandran plot (fig. 5.8).

For further details regarding helicity especially in this puzzling state, fig. 5.13 depicts the fractions of three types of H-Bonds versus T for S_{16} and Q_{16}, i.e. the average numbers of these H-Bond types divided by the maximum possible numbers of such H-Bonds. The three types are defined by the positions of the bond partners, with α signifying $\mathrm{NH}_i \cdots \mathrm{CO}_{i-4}$ bonds as they occur in α-helices and γ standing for

$NH_i \cdots CO_{i+4}$ bonds. An unnamed third type, $NH_i \cdots CO_{i-6}$, is found commonly as a defect in γ-helices especially of polyA chains (cf. fig. 4.10 F). The respective maximum possible numbers of these types of H-Bonds are 12, 12, and 10. α H-Bonds are shown as green lines, γ H-Bonds in red and $(i, i-6)$ H-Bonds in blue. The line types are the same as in the previous graphs: solid lines for PRIME20n, long dashes for PRIME20s and short dashes for PRIME20.

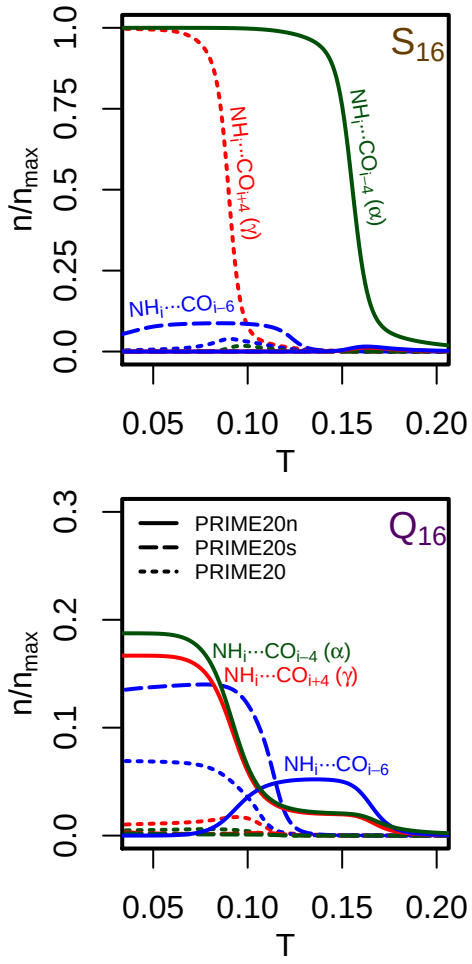

Figure 5.13 – Fraction of $(i, i-4)$ (α-helical), $(i, i+4)$ (γ-helical) and $(i, i-6)$ H-Bonds of polyS and polyQ versus temperature in PRIME20, PRIME20s and PRIME20n.

The top graph of fig. 5.13, depicting H-Bonds of S_{16}, unsurprisingly contains two prominent lines representing the γ-type H-Bonds in PRIME20 and the α-type H-Bonds in PRIME20n. Both increase to a fraction of 1 at the same temperatures at which the respective $n_{\mathrm{HB}}(T)$ shoot up as well. This observation confirms the native states of polyS in both models to be helical. A very small contribution of $(i, i-6)$ defects in the PRIME20 γ-helix can be observed around the transition temperature, but it vanishes again at lower T. In PRIME20s, the low-T state contains an average of 0.9 $(i, i-6)$ H-Bonds while α- and γ-type H-Bonds do not occur at all. This observation comes as no surprise because α- and γ-turns have been identified to be inaccessible in PRIME20s.

In the bottom graph, which depicts the same H-bond types for Q_{16}, the α and γ lines are less prominent because these helices are not formed in PRIME20/PRIME20n. The yet-unknown low-T state in PRIME20n is characterised by a noticeable contribution of both bond types, each making up 2 of the 8 H-Bonds in the configuration. It is clearly not a helical state; instead the occurrence of these H-Bonds suggests a structure with two rather tight turns, each turn stabilised by one α and one γ H-Bond. In the other model variants, most of these H-Bond types play a largely negligible role, so this curious structure occurs exclusively in PRIME20n. A somewhat prominent fraction of $(i, i-6)$ H-Bonds is observed in the intermediate PRIME20n state as well as in the low-T state of PRIME20s, but this fraction only corresponds to a single such H-Bond in an entire conformation, so it can hardly be considered a defining element of the state.

5.4 Simulation results: Complete structures

The preceding section treated the temperature-dependent behaviour of polyS and polyQ in the three PRIME20 variants. In this section, the low-temperature states will be identified using the more detailed information of H-Bond contact matrices.

Each of the figures 5.14 to 5.20 refers to either S_{16} or Q_{16} in one of the three established model variants. Figures 5.21 and 5.22 depict data of Q_{10} in PRIME20n and of A_{16} in PRIME20s. In every figure, the top graph shows the distance distribution $p(d, T)$, using the format first seen in fig. 4.8, i.e. a color scale ranging from white ($p \approx 0$) via turquoise to red ($p \approx 0.1$ in a bin of width 0.1 Å). The bottom picture is an HB matrix at low temperature, and in some cases another HB matrix of an intermediate state, i.e. at a temperature between two transitions, is added. Attached to each HB matrix is a typical configuration snapshot.

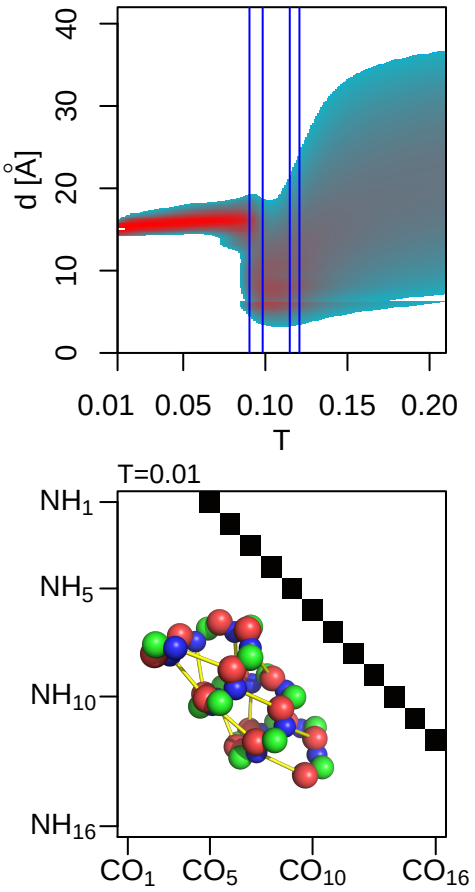

Figure 5.14 – Terminal side chain distribution function $p(d,T)$, low-temperature HB matrix and corresponding γ-helical configuration snapshot of S_{16} in PRIME20.

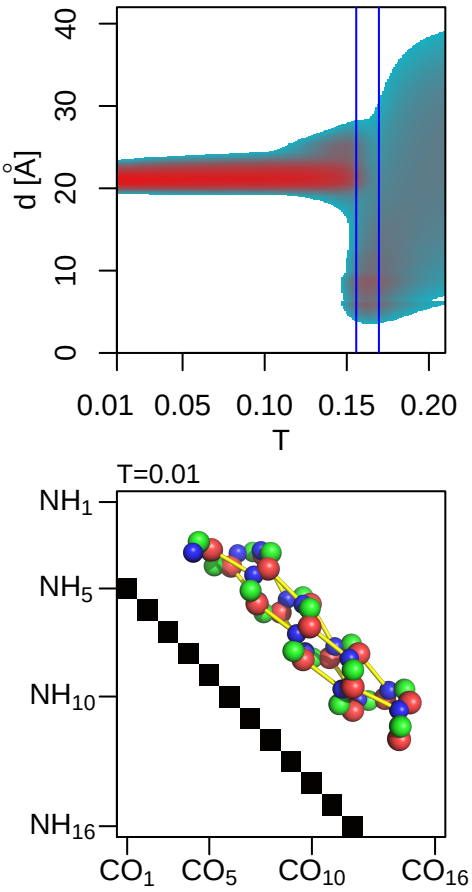

Figure 5.15 – Terminal side chain distribution function $p(d,T)$, low-temperature HB matrix and corresponding α-helical configuration snapshot of S_{16} in PRIME20n.

The first system is S_{16} in the PRIME20 model (fig. 5.14). Its configurations were discussed along figures 4.8 and 4.10 already. In summary of the findings there, microcanonical analysis identified four transition temperatures $T = \{0.090, 0.098, 0.114, 0.121\}$, of which two were assigned to coil-globule and globule-helix transitions. The configuration at lowest T is an unperturbed γ-helix, as indicated by the $(i, i+4)$ diagonal in the HB matrix. This helix can be right- or left-handed and both versions are observed with no apparent preference for either version.

Figure 5.15 shows S_{16} in the PRIME20n model. Contrary to the single heat capacity maximum seen in fig. 5.11, the $p(d,T)$ image clearly shows the existence of a globule state and accordingly contains two microcanonical transitions at $T = \{0.156, 0.170\}$. The distance distribution in the globule state is similar to its PRIME20 counterpart, with a maximum somewhere around the square-well diameter. Compared to PRIME20 however, it extends to higher d, anticipating the folded state, which features a narrow distribution around 21 Å opposed to 15 Å in PRIME20.

The discussion of number and type of H-Bonds has already confirmed the ground state of S_{16} in PRIME20n to be an α-helix, and as expected the HB matrix consists solely of α-helical contacts, yielding a regular structure as depicted in the snapshot. The helix is narrower and longer than the γ-helix, which causes the aforementioned difference in terminal side chain distances.

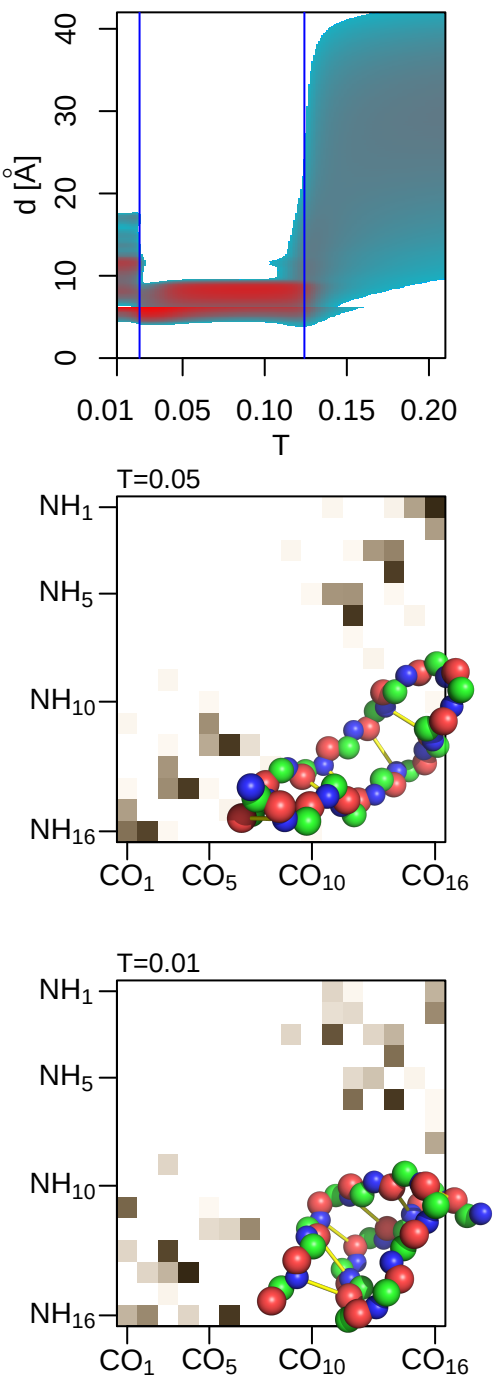

Figure 5.16 – Terminal side chain distribution function $p(d, T)$, HB matrices at two temperatures and corresponding configuration snapshots of S_{16} in PRIME20s.

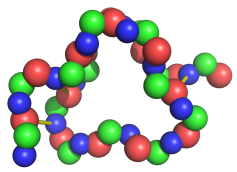

Figure 5.17 – Low-T configuration of fig. 5.16 viewed from a different angle.

While the PRIME20 and PRIME20n results are somewhat similar (except for the different helix types), the situation in PRIME20s (fig. 5.16) differs. As already seen, the behaviour of S_{16} in this model variant is dominated by a collapse transition at $T = 0.124$. A second transition at $T = 0.024$ has not been mentioned yet. At such a low temperature (about 45 K according to the results shown in section 6.1), it is irrelevant in a biological context, but interesting nonetheless to gain insight into the configurations which the model can produce.

The state between the transitions, which is therefore to be regarded as intermediate rather than the ground state, is characterised by an ascending diagonal pattern in its H-Bond matrix. During the discussion of Q_{16} in PRIME20 (section 4.4), this pattern was found to be typical for a hairpin configuration. Contrary the situation there, the population here is mostly limited to the main diagonal and first side diagonals. A hairpin on the more distant side diagonals would contain a rather tight turn between residues $(i, i \pm 4)$. Based on the observation from the Ramachandran plots that α- and γ-turns are illegal in PRIME20s, it is conceivable that all turns with these H-Bonds may be difficult to form (but not impossible, as they do occur in the HB matrix of Q_{16}, fig. 5.19). The absence of distant side diagonals in the HB matrix supports this assumption and thereby confirms an overall lower flexibility of PRIME20s, not only in the context of helices

Below $T = 0.024$, the HB matrix layout changes. While the hairpin is still visible, a new signature dominates. It is characterised by H-Bonds on the diagonal between $NH_{11} \cdots CO_1$ and $NH_{16} \cdots CO_6$ as well as the isolated spots $NH_3 \cdots CO_{11}$ and $NH_6 \cdots CO_{14}$. The mirror image of these H-Bonds is populated as well, albeit weakly. A configuration with the described H-Bond pattern is attached to the figure. In this view, it has the appearance of a large loop. Viewed along the main axis (fig. 5.17), it exhibits a structure of three narrow turns which are not stabilised by H-Bonds and instead connect into a "clover" shape. The energy of this structure is equal to that of the bent hairpin depicted in fig. 5.16, but lower than that of straighter hairpins, which causes this structure to dominate at lower temperatures. By adding further amino acids, the clover loop could conceivably be extended to a helix with a higher number of H-Bonds (and correspondingly lower energy) than a hairpin.

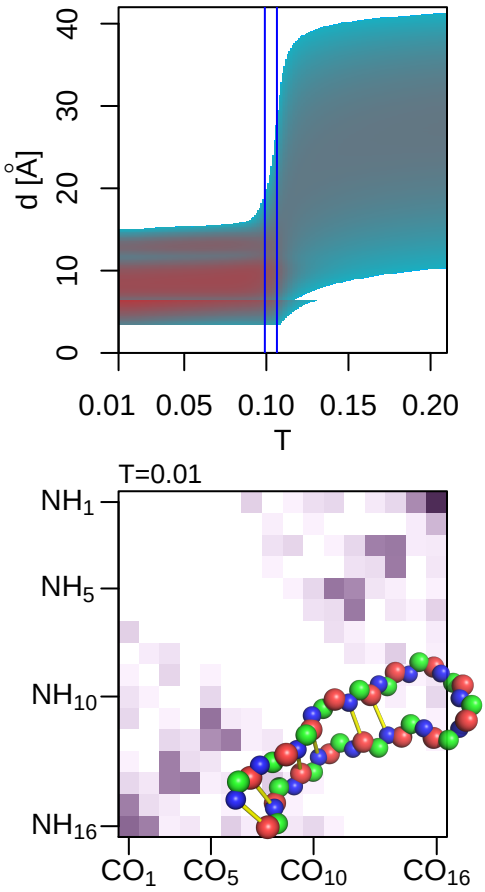

Figure 5.18 – Terminal side chain distribution function $p(d, T)$, low-temperature HB matrix and corresponding hairpin configuration snapshot of Q_{16} in PRIME20.

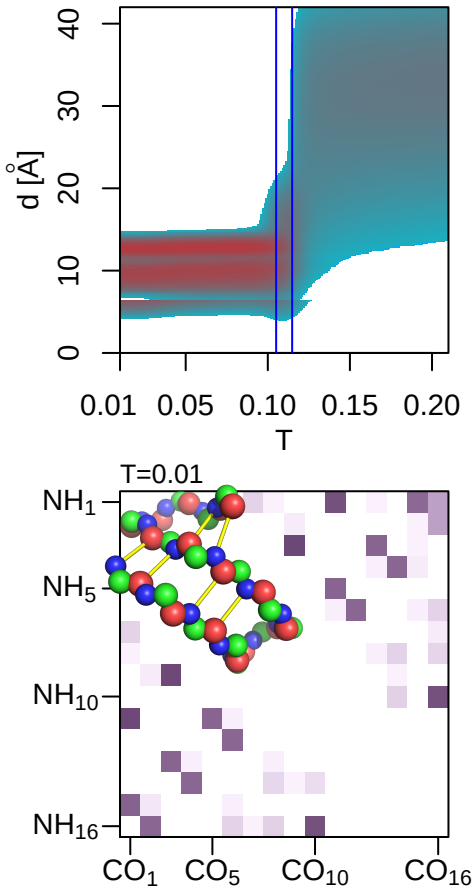

Figure 5.19 – Terminal side chain distribution function $p(d, T)$, low-temperature HB matrix and corresponding double-hairpin configuration snapshot of Q_{16} in PRIME20s.

Turning to Q_{16}, figure 5.18 shows its behaviour in the PRIME20 model. To recapitulate the discussion from section 4.4, the peptide collapses from a random coil to a hairpin – either directly or through a very narrow globule state, as the two microcanonical transitions at $T = \{0.100, 0.106\}$ indicate. Unlike the case of S_{16}–PRIME20n shown in a previous paragraph, a distinct signature of the globule state cannot be identified here. The HB matrix at low T consists of the well-known hairpin signature as well as several weakly populated cells. These may hint at double-hairpin configurations, which will be seen and discussed in the subsequent PRIME20s/n cases.

The PRIME20s version of Q_{16} is depicted in figure 5.19. Side to side with fig. 5.18, both $p(d, T)$ graphs look very similar. In the random coil region, the PRIME20s chain is overall more extended than the PRIME20 chain, which comes as no surprise given the lower flexibility in PRIME20s. The transition temperatures, $T = \{0.105, 0.115\}$, are comparable to the PRIME20 values, and both low-T states feature the same three maxima at $d \approx \{6.4, 10, 13\}$ Å, albeit with different intensities.

The HB matrix of Q_{16} in PRIME20s features the same hairpin signature as in PRIME20. The main diagonal is not strongly populated, probably just due to insufficient visitation of low-energy states. A second prominent signature can be seen, consisting of cells on two of the shorter diagonals. The attached configuration snapshot of a double hairpin belongs to this signature, which – as noted before – is seen faintly in the PRIME20 HB matrix as well.

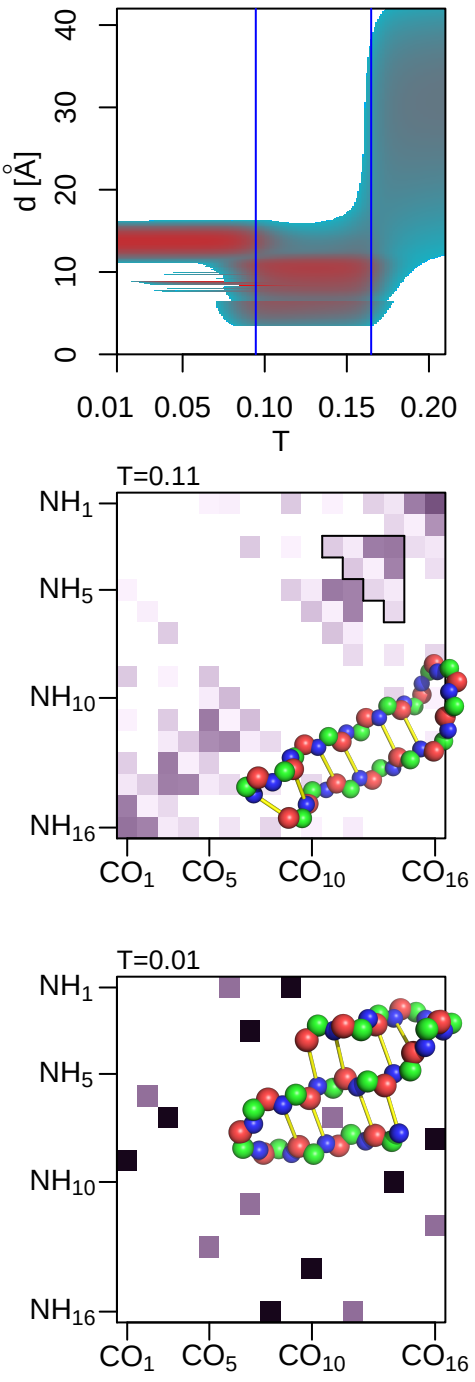

Figure 5.20 – Terminal side chain distribution function $p(d,T)$, HB matrices at two temperatures and corresponding single and double hairpin configuration snapshots of Q_{16} in PRIME20n.

The distance between the side chains R_1 and R_{16} in the snapshot is 12.5 Å, indicating that the 13 Å feature of $p(d,T)$ belongs to this type of configuration. Compared to PRIME20, the signature in the HB matrix is more prominent, which relates to the higher intensity of this $p(d,T)$ maximum. However, there is no obvious reason why either configuration should be preferred to the other. With six H-Bonds each, they are energetically identical. The large number of double hairpins here, like the lack of single hairpins on the main diagonal, may just be caused by insufficient sampling of configuration space.

In PRIME20n (fig. 5.20), the double hairpin state has a lower energy than the single hairpin, which causes a change in $p(d,T)$ as well. Contrary to the other models, two distinct transitions can be observed at temperatures $T = \{0.095, 0.165\}$. The transition at higher temperature is the familiar collapse from a random coil to a hairpin with its typical characteristics seen in earlier plots too. Unlike PRIME20s, PRIME20n has sufficiently easy access to $(i, i \pm 4)$ H-Bonds, thus allowing the full range of shifted hairpins which has been described for PRIME20 in the previous chapter. In comparison to PRIME20, a small detail becomes visible too, which appears to be a systematic behaviour: within each of the six triangular spots (of which one is marked in the plot for clarity), the "hypotenuse", i.e. the set of cells indicating a configuration which is derived from a Q_{13} hairpin, is preferred over the Q_{14} line (the set of cells neighbouring the hypotenuse). This indicates that the longer loose ends allow a slightly lower potential energy due to additional side chain contacts, creating a hierarchy between the hairpin types which was not as clearly recognisable in PRIME20.

At $T < 0.095$, the single hairpin configuration disappears in favour of a double-turn structure similar to the one seen in PRIME20s. Contrary to the double hairpins in PRIME20s and PRIME20 however, this configuration contains eight H-Bonds, allowing a lower energy than in the single hairpin and causing the clear transition between two folded states. This is of course the same transition as the "unidentified" feature in the previous sections (figures 5.11-5.13 and table 5.3).

The HB matrix also features a second, weaker signature consisting of seven H-Bonds. Their positions on three separate diagonals suggests a kind of three-turn configuration in which both ends of a central hairpin bend back onto itself. Unfortunately, no snapshot of this configuration could be obtained because the favourable conformation with two turns and eight H-Bonds dominates the low-T behaviour.

The treatment of S_{16} and Q_{16} in the three PRIME20 variants largely confirms the expectations which were formulated in the Ramachandran plot analysis. Polyserine folds into helical ground states in PRIME20 and PRIME20n, but in PRIME20s both are prevented, either by the added side chain squeeze parameters (γ-helix) or by the missing backbone squeeze factors (α-helix), hence the configurations of lowest energy are hairpins and the clover shape.

In the case of polyglutamine, both helix types are hindered by interactions between side chain and backbone beads of neighbouring turns. Therefore, polyQ forms hairpins even at low temperature in all three models. The main difference between them is the ability of PRIME20n to form a double hairpin with more H-Bonds, producing a low-temperature state which does not exist in PRIME20 and PRIME20s.

The $N = 10$ chains were not treated explicitly here because their behaviour is largely the same as that of $N = 16$. An interesting exception is Q_{10} in PRIME20n, shown in fig. 5.21. At high to medium temperatures, the peptide behaves as expected, folding from a random coil into the familiar hairpin at $T = 0.146$. However, a second transition is observed at $T = 0.064$. The distance distribution below this temperature differs from that seen for Q_{16} in this model, and the double hairpin observed there cannot be formed by this short chain. Instead, the $p(d, T)$ maximum jumps from 10 Å to 15 Å and (without a further transition signature) back to 8 Å at lower T.

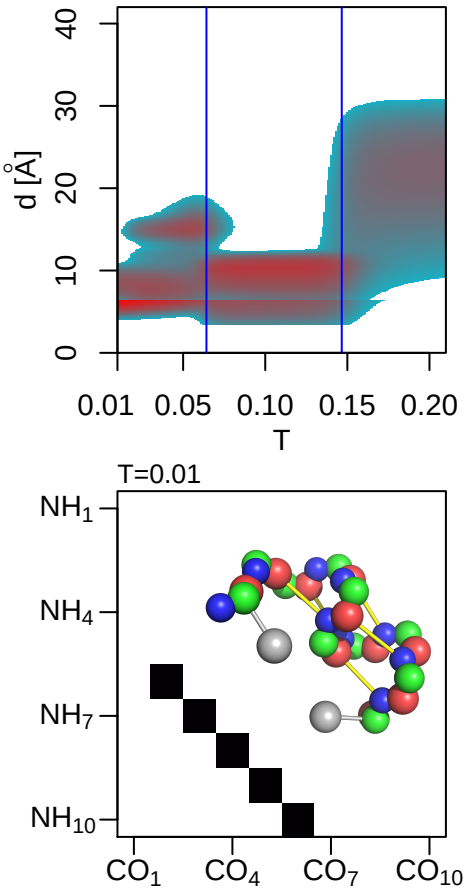

Figure 5.21 – Terminal side chain distribution function $p(d, T)$, low-temperature HB matrix and corresponding α-helical configuration snapshot of Q_{10} in PRIME20n. The side chain beads R_1 and R_{10} are depicted by grey spheres and sticks.

As the contact matrix indicates, the low-T structure is an α-helix fragment. Five out of six α-type H-Bonds are closed with a probability of 100%, but the $NH_5 \cdots CO_1$ contact does not occur at all. This behaviour shows that forming one α-helix turn is possible for polyQ (matching experimental observations according to which short polyQ segments tend to form α-turns [118, 147]), but as the helix grows longer, accommodation of the large side chains requires deformations which seem to cause an essential break at the 10th amino acid of a helix fragment.

As an experimental side note, even though polyQ is usually disordered, single Glutamine residues have been found to stabilise α-helical configurations [118, 147] and the somewhat related PBLG (poly-γ-benzyl-L-Glutamate) has been known to be α-helical for a very long time [55, 160], forming stable helices even far above room temperature [198]. PolyQ itself has been found to at least be capable of forming α-helix segments under certain conditions [25, 99]. Based on this literature, the observation of an α-helix in polyQ does not seem too surprising and the helices should even be much more common. However, polyQ in PRIME20n does not meet this expectation due to a repulsion of the spherical side chains. The real side chains are long and flat and can be arranged along the helix more easily. A more accurate side chain representation consisting of multiple beads has been attempted elsewhere [72, 121] and it has been suggested as a possible extension to PRIME20 too [210, sec. 5.1a], but such representations contradict the postulate of computational simplicity, which after all is the reason to use coarse-grained models.

As a second side note, the terminal side chains in a regular α-helix of length 10 have a well-defined distance. For S_{10}, this distance is 13.6 Å; for Q_{10} it might be slightly larger due to the longer C_α–R bond. The spot at $T \approx 0.05$ and $d \approx 15$ Å corresponds to such a configuration. At lowest T, the distances decrease, which appears contradictory because the lack of a transition signature suggests that the system remains in the helical state. An explanation for this behaviour is found in the configuration snapshot: the free C-terminus bends back towards the helix to maximise the number of side chain contacts. Among the contacts formed this way is the R_1–R_{10} interaction (side chains shown as grey beads), here with a distance of 4.6 Å. Configurations like this explain how the low-d feature in the $p(d, T)$ plot comes to be without requiring a major structural rearrangement.

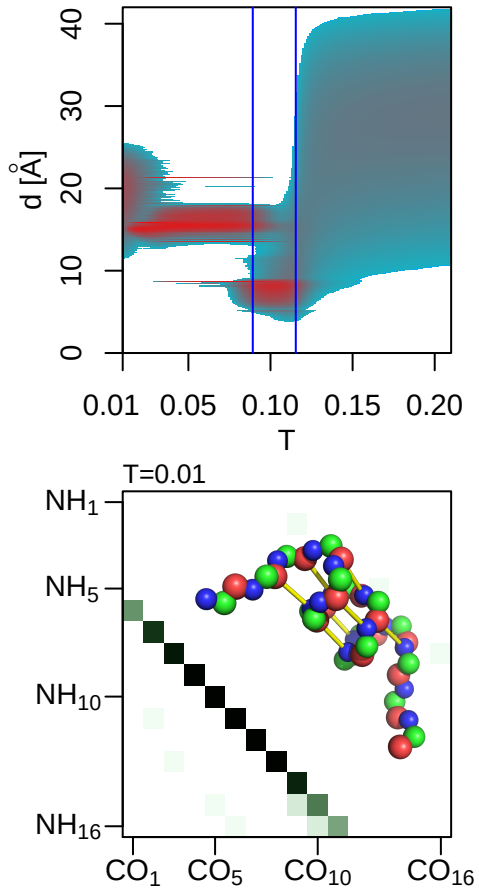

Figure 5.22 – Terminal side chain distribution function $p(d, T)$, low-temperature HB matrix and corresponding π-helical configuration snapshot of A_{16} in PRIME20s.

Finally, one surprising structure occurred in the simulations of A_{16} using the PRIME20s model. As shown in the $p(d, T)$ part of fig. 5.22, the chain has two transition signatures at $T = \{0.089, 0.115\}$. Parallel to the behaviour of S_{16} (see fig. 5.16), the transitions mark a collapse of the random coil into a hairpin and a refolding event from the hairpin to the clover-type configuration. At very low temperatures ($T \lesssim 0.02$) however, the distance distribution jumps from its "clover" maximum at $d \approx 15$ Å to a broader distribution around 20 Å. This jump could just mean a rearrangement of a loose end like in the Q_{10} α-helix discussed before, but the low-T HB matrix exposes the jump as a transition between distinct configurations. A transition signature is missing, probably because not all simulation runs reached this state.

The HB matrix characterises this low-T state as a π-helix, defined by regular $NH_i \cdots CO_{i-5}$ contacts. The helix does not extend over the entire chain, but all possible contacts are formed with nonzero probability, suggesting that the same helix could be completed as a regular structure (unlike the Q_{10} α-helix just seen, which cannot extend beyond five consecutive H-Bonds). The rather broad d distribution is caused by the flexible loose ends of those incomplete π-helices.

Although low-T structures were argued to be of little interest and the PRIME20s model itself is just a purely hypothetical hybrid of PRIME20 and PRIME20n, the very occurrence of such a structure is interesting because, as interpreted from the respective Ramachandran plots in the first section of this chapter, it should not be legal at all. Both the α_R- and π_R-helix regions have been ruled out in PRIME20s by the central oval of the CO_i–NH_{i+2} repulsion. This observation is reiterated in fig. 5.23, the PRIME20s Ramachandran plot of polyA, in which the π_R region is clearly inaccessible due several overlaps.

The red circles in this Ramachandran plot mark the helical residues of the snapshot in fig. 5.22. Clearly, all of them are "illegal" and none of them lie exactly within the narrow π_R region. As stated in the first section, expectations regarding the availability of structures can be drawn from these plots, but they do not represent rigorous rules. Further configurations can be made available by deformation of bond angles and of the peptide bond dihedral ω, and the A_{16} π-helix is an example of such a configuration.

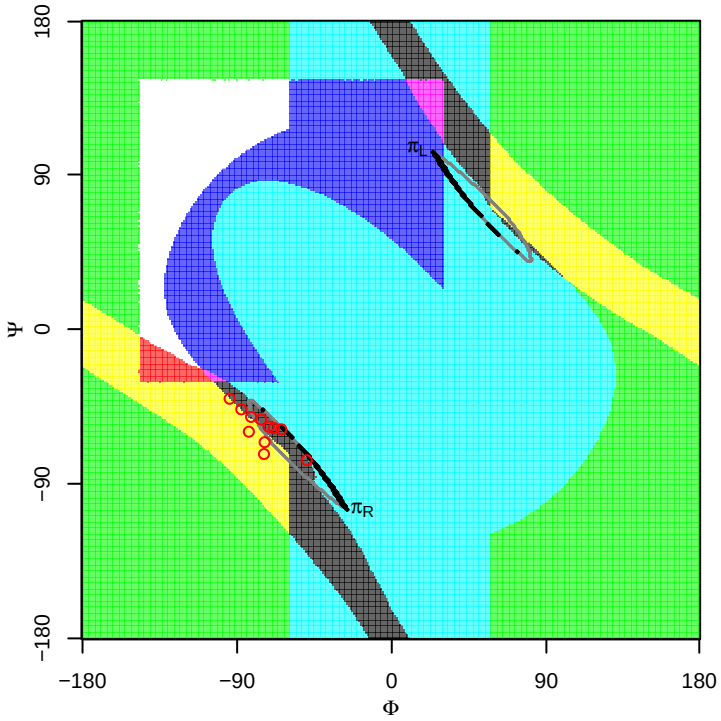

Figure 5.23 – Ramachandran plot of A_{16} in PRIME20s. The (Φ, Ψ) pairs selected by the configuration in fig. 5.22 are marked by red circles and the π-helical regions outlined in the same way as in the previous Ramachandran plots.

Notwithstanding this observation, α- or γ-helices did not occur in simulations and PRIME20s most likely remains unable to form them due to the geometric constraints. PRIME20 and PRIME20n are more similar to each other regarding the range of available energies (which relates to these helices), but the higher flexibility of the PRIME20n backbone shows its effect in the formation of double hairpins and, most importantly, of α- rather than γ-helices, making this model variant the one that should be used in future simulations.

Chapter 6

Experiment and simulation

The results collected in the previous two chapters are entirely theoretical. The present chapter treats simulations of polyglutamine-based peptides as they are used in spectroscopy experiments and compares the results to experimental data. In this collaborative project, Peter Enke and Michael Schleeger investigated the formation of "loops" in polyglutamines and polyserines experimentally, i.e. of configurations in which the ends of these peptides are in contact. The corresponding SAMC simulations are described in the following sections.

Polyglutamine sequences occur naturally in several proteins. In some cases, these sequences are extended and cause neurodegenerative diseases including Huntington's chorea [4, 201, 231]. Amyloid aggregates of polyglutamine have been found in brain tissue of persons afflicted with these diseases. A handful of similar polyalanine diseases are known as well, as described in more detail in the introductory chapter. In all of these cases, the disease mechanism is not entirely understood yet, which motivates research on the structure formation of polyglutamines and other homopolymeric peptides.

The experimental methods in this collaboration are Förster Resonance Energy Transfer (FRET) and Triplet-Triplet Energy Transfer (TTET) spectroscopy, applied by Peter Enke and Michael Schleeger. The methods differ in their physical background and in the obtainable information, but from a simplified simulation point of view they are practically identical: in both cases, two chromophores are grafted to the peptide, small molecules used to absorb or emit photons with well-defined wavelengths. One chromophore is excited by a matching laser pulse, then the energy is transferred to the second one which finally emits a photon at its own wavelength. Because the energy transfer rate depends on the distance between the chromophores, the emission intensity provides insight into the configurations of the peptide.

The main comparable result between FRET and the simulations is the equilibrium distance distribution between the chromophores. TTET, which requires a direct contact between the chromophores for energy transfer, yields a probability of contact formation. Further dynamic data can be extracted from both experiments, but this type of information is not accessible by means of SAMC simulation. Comparable simulations using Molecular Dynamics have been performed by Svetlana Pylaeva within the framework of this collaboration as well and parts of the results can be found in a joint paper by Pylaeva et al. [166].

As detailed in the model chapter, the chromophores are best represented as tryptophans in PRIME20, even though their chemical structures differ. Due to its large size and its hydrophobic properties, the side chain interactions of tryptophan in the model are considerably stronger than those of glutamine, serine or alanine. Hence, the chromophores can easily be expected to affect structure formation, like they have been shown to do experimentally elsewhere [209]. The aim of the present chapter is therefore to characterise how large the influence of the tryptophans in this set-up is.

Furthermore, the water solubility of the peptides needs to be enhanced for the FRET and TTET experiments, which is achieved by adding polar residues to their C-termini. Here, a tail of five amino

acids (Ser-Arg-Ser-Arg-Gly) is used, which is believed not to interact with the main chain in the relevant temperature range, but direct evidence for this claim is lacking. In implicit solvent simulation, the tail can be left out of the chain, making it possible to investigate its influence on structure formation.

Various peptides have been and are being investigated with FRET/TTET; corresponding simulations were performed on polyglutamines and polyserines of length 9 and 14. Additionally, paralleling the previous chapters, polyalanines of the same lengths were simulated as well. They are more interesting than in the other chapters because the serine residues in the solubility tail will interact with a polyserine chain differently than with a polyalanine chain. For each of the peptides, three modifications were considered: 1) the peptide itself, 2) the peptide with a tryptophan added to each end, and 3) the peptide with tryptophans and the solubility tail attached to the C-terminus. In the first case, the chain is elongated by two residues in order to better compare the distance distributions to those of the tryptophans. As an example the Q_{14}-derived systems have the amino acid sequences Q_{16}, $WQ_{14}W$ and $WQ_{14}WSRSRG$. For better readability, they will be called Q_{16}, $Q_{14}c$ (+**c**hromophores) and $Q_{14}ct$ (+**c**hromophores+**t**ail) for the remainder of the chapter. The same terminology is applied to polyS and polyA.

The essence of chapter 5, summarised briefly, is that the PRIME20n model produces more realistic conformations than PRIME20. However, the present chapter is based on simulations which were performed prior to those findings. At first glance this could mean that the PRIME20 results are obsolete and all simulations need to be redone using PRIME20n. But if the following arguments hold, the results should still be significant, if the model weaknesses are kept in mind.

The most noticeable differences between PRIME20n and PRIME20 discussed in the previous chapter comprise three issues:

1. much higher transition temperatures in PRIME20n than in PRIME20,
2. polyA/polyS forming α-helices in PRIME20n opposed to γ-helices in PRIME20,
3. Q_{16} forming a stable double hairpin at very low temperatures in PRIME20n.

Regarding the first issue, the reduced temperatures (T) from simulation need to be scaled to physical temperatures ($T'\,[K]$) for the comparison to experiments. This scaling, to be presented in the following section, will effectively be based on the collapse temperature, so the different collapse peaks in PRIME20 and PRIME20n would be mapped to the same physical value. Thus, the physical transition temperatures themselves do not depend on the choice of model. Instead, what depends on this choice is the ratio T'/T, which in turn relates to the effective hydrogen bond strength ϵ_{HB}. This quantity will be discussed for both model variants.

The second issue, the type of helix in the ground state, is less straightforward. Both types of helix, α and γ, contain up to $N-4$ H-Bonds, so their energies are equal barring side chain contributions. This means that at least at a qualitative level, the prevalence of helical or nonhelical configurations should not depend on the type of helix and thus on the choice of model. Judging by the configurations observed in simulation, even the available side chain energies are equal between both helix types, suggesting a quantitative comparability between both models regarding their helix formation behaviour. However, this remains a purely speculative argument with no evidence supporting or rejecting it.

Finally, the double hairpin of Q_{16} dominates the lowest temperatures in PRIME20n. In this chapter however, the focus lies on higher temperatures, around the transition between the random coil and (single) hairpin or globule states. This temperature range is not affected by the existence of further low-temperature states.

Both the second and third issues can be regarded as symptoms of the different backbone flexibilities in PRIME20 and PRIME20n. This geometric difference affects all temperatures including the random coil state. A comparison of the $p(d,T)$ graphs in figures 5.14, 5.15 (polyS), 5.18 and 5.20 (polyQ) indicates a notable difference because the average chromophore distance in the random coil state tends

to be longer in PRIME20n than in PRIME20. Further such nuances of the peptide behaviour might differ, but should at least not change the qualitative interpretation of results.

The remainder of the chapter is organised as follows: in the first section, the process of scaling reduced temperatures to physical temperatures is discussed. Using this physical scale, thermodynamics and folding behaviour of the aforementioned modifications of Q_{14}, S_{14}, A_{14} are studied at biological temperatures as well as in a "frozen" state below 0° C. As mentioned earlier, peptides with $N = 9$ amino acids were investigated too by experiment as well as simulation, but they are left out of this analysis as their behaviour is largely the same as that of the $N = 14$ cases and the main focus lies on $N = 14$.

6.1 Temperature scales

A prerequisite to compare simulated and experimental results is a temperature scale. Implicit solvent simulations often use reduced, dimensionless energies and temperatures, which need to be scaled to physical quantities. This scaling procedure depends on the effective strength ϵ_{HB} of a backbone H-Bond in the system because the physical energy is $E' = \epsilon_{\mathrm{HB}} E$ and the temperature $T' = {}^{\epsilon_{\mathrm{HB}} T}/_{k_B}$. ($E$ and T denote the reduced quantities and E' and T' the physical ones, as defined in the methods chapter.) Knowledge of ϵ_{HB} would therefore directly provide a sensible temperature scale (also expressed by a conversion ratio T'/T).

Unfortunately, although the question is well-defined this way, its answer, i.e. a value of ϵ_{HB}, is difficult to come by. Estimates in literature vary wildly between +12 kcal/mol [120] and about −5 kcal/mol [228], if the competition between H-Bonds within the backbone and H-Bonds between backbone and solvent molecules is taken into account. (Note that in PRIME20 the H-Bond energy is -ϵ_{HB}, hence attractive ϵ_{HB} values are positive in this notation.) Most values vary around 3-6 kcal/mol [66, 69, 81, 90, 132, 171, 181], depending on the method of investigation, but also on further parameters like temperature and solvent quality. In models related to PRIME20, values like 3.8 kcal/mol [187] or 5 kcal/mol [50] have been used in the past. (Interestingly, both cite the same source [81], which itself provides multiple answers to this question.) In publications using PRIME/PRIME20, T'/T conversion ratios of 3300 K [205, calculated from Table II] or 2288 K [219] can be found for polyA-based peptides, which translate to ϵ_{HB} values of 6.6 kcal/mol and 4.5 kcal/mol respectively.

Even though this collection is far from comprehensive, it is already too large and varied to answer the question in an unambiguous way. Picking an arbitrary value from it cannot be a satisfying solution. Instead, a juxtaposition of simulation data and available experimental results will be performed to produce new ϵ_{HB} values. This method essentially adds even more numbers to the above collection, but they will arguably be correct for the specific set-up here.

FRET/TTET results of polyQ and polyS are available and will be discussed subsequently. For polyA, such data do not exist, so another approach is required. In this case, the well-documented folding behaviour from literature will serve as a gauge. PolyA chains have been found to be in the helix state at biological temperatures, so the folding transition must be higher. Different sources report transition temperatures somewhere around 350 K [172, 176, 185, 195] or 450 K [75, 146, 150, 188, 221]. The discrepancy may be explained by the inclusion of polar residues in the chain, a necessary tool used to enhance solubility of the polyalanine chain. The authors who used this tool observed the transition around 350 K while the higher temperatures apply to pure polyalanines in implicit solvent simulation. The latter results are therefore better suited for the PRIME20 scale here.

In the most recent of the comparable polyA sources, Wei et al. find a helix-coil transition of A_{10} at $T' = 462$ K. In PRIME20, the transition is split into a helix-globule part at $T = 0.093$ and a globule-coil part at $T = 0.108$. The most reasonable, albeit unphysical, course of action seems to be assigning the 462 K to their average ($T = 0.100$). This results in a T'/T ratio of 4620 K and an ϵ_{HB} value of 9.2 kcal/mol.

Compared to the most cited sources in the aforementioned collection reporting 3-6 kcal/mol, this value appears unreasonably high. However the method is justified by a quick estimate using PRIME20n: in this model, the A_{10} folding transition lies at $T = 0.152$, which translates to $T'/T = 3040\,\mathrm{K}$ and $\epsilon_{HB} = 6\,\mathrm{kcal/mol}$. This value is within the expected range, and furthermore T'/T resembles the values around 3300 K which were applied in the first PRIME papers [205]. Thus, the high values can be considered artefacts of PRIME20 and the procedure itself, applied to the correct model variant, produces reasonable numbers.

As a side note, in a later PRIME20 (i.e. PRIME20n) publication Wang et al. present a T'/T ratio of 2288 K [219]. This seems contradictory at first glance, but they used a solubilised form of polyA, which, as discussed before, folds into a helix around 350 K. With this transition temperature, ϵ_{HB} calculates to 6 kcal/mol, matching the PRIME20n value of A_{10} to within one percent.

While polyA is known to be helical at room temperature, polyS is usually reported to be disordered [27, 77, 82, 104]. This means that the transition temperatures of polyS must lie much lower than those of polyA. Since the reduced temperatures do not differ much, this discrepancy will be contained in ϵ_{HB}. PRIME20 does not take side chain hydrogen bonding into account explicitly, so the alanine ($-CH_3$) and serine ($-CH_2OH$) side chains are treated very similarly. In reality however, the serine side chain differs from alanine by its capability of forming hydrogen bonds with the backbone or with solvent water molecules. These hydrogen bonds affect the competition between backbone and solvent H-Bonds, which is essentially expressed by ϵ_{HB}. For this reason, ϵ_{HB} must be different between polyA and polyS, leading to the different transition temperatures. The glutamine side chain forms H-Bonds as well, but due to the different number and nature of polar groups and the overall side chain length the resulting ϵ_{HB} should differ as well.

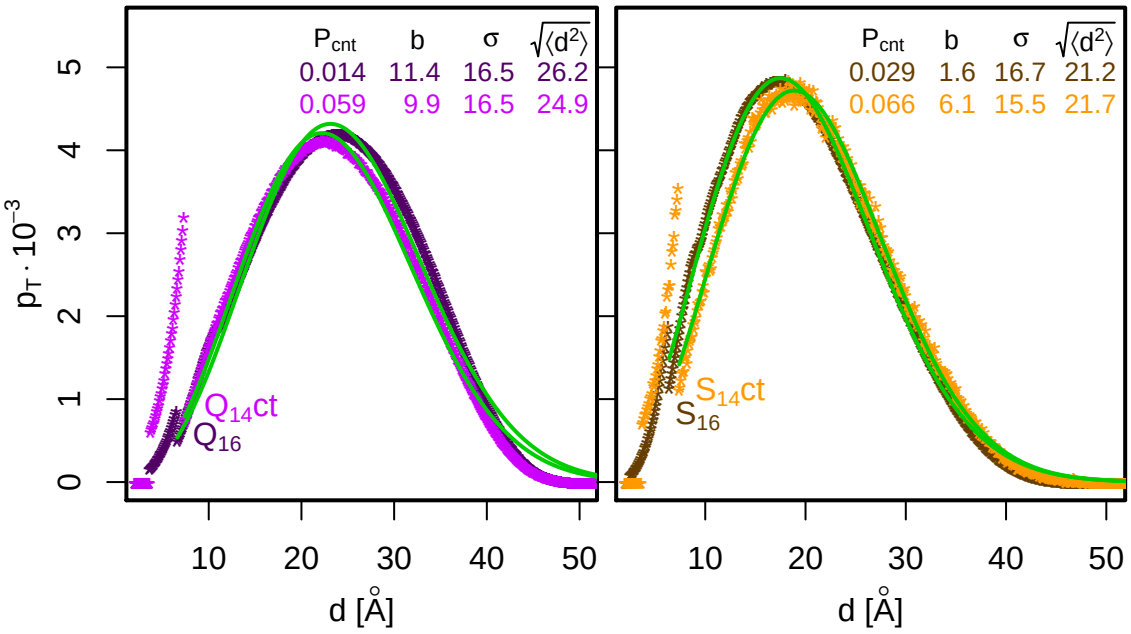

Figure 6.1 – Distance distributions of Q_{16}, Q_{14}ct at $T = 0.135$ and the corresponding polyS chains at $T = 0.16$. The Edwards model fits (see eq. 3.29 and fig. 4.9) are shown in green.

To obtain these, FRET/TTET results are compared to those from simulation. The chromophore distance distributions $p(d, T)$ were shown multiple times in the previous chapters, and the same distributions were calculated from FRET decay curves by Peter Enke using the Edwards polymer model (eq. (3.29)), which has been found to be the most suitable for these experiments [61]. The same fit was performed

on the simulation results as an example in the first chapter (fig. 4.9), yielding two fit parameters b and σ. Figure 6.1 now shows the chromophore distance distributions of Q_{16}, Q_{14}ct (left side), S_{16} and S_{14}ct (right side) as points and the respective Edwards fits as green lines. The selected temperatures are 0.135 for the polyQ plots and 0.16 for polyS. These two temperatures will turn out to be assigned to 300 K in the course of this section.

At these temperatures, the chains are in their random coil state, which (ignoring the low-d section) produces bell-shaped distributions as premised by the attempt to fit them with the Edwards model. The fit curves themselves follow a similar bell shape, albeit with some deviations. For one, in the polyQ graph a population of already collapsed configurations with $d \approx 13$ Å causes a small bulge which the fit cannot include. If the temperature is decreased, approaching the collapse temperature, this population grows in magnitude and ultimately renders the fit unable to converge. In the polyS graphs at $T = 0.16$, the same population is yet negligible, but a similar behaviour can be observed at lower temperatures.

The second deviation between fit and simulation data is the tail at large d, which – again more visibly in polyQ than in polyS – is overrepresented by the fit. The Edwards model does not regard attractive interactions. These interactions slightly reduce the end-to-end distance, which leads to the discrepancy at large d.

Lastly, as discussed earlier, the attraction between side chains causes a jump in $p_T(d)$ at the respective square well diameter. Below this distance, the chromophores are in contact, which is essential for the TTET experiment. One of the results obtainable by means of TTET is the probability P_{cnt} for this contact to be formed at a given temperature. The integral of $p_T(d)$ or, due to the discretisation, the sum $\sum_{d_i=0}^{d_{\mathrm{SW}}} p_T(d_i)$ is equal to P_{cnt} and can be used for temperature scaling as well.

In the top right corners of both graphs in fig. 6.1, the fit parameters b and σ (in Å) are noted as well as P_{cnt} and the root mean average squared distance, $\sqrt{\langle d^2 \rangle} = \left(\int_{d_{\mathrm{SW}}}^{d_{\mathrm{max}}} d^2 \tilde{p}_T(d) \mathrm{d}d \right)^{1/2}$ (in Å), where $\tilde{p}_T(d)$ is the fit function. This distance, related to the radius of gyration, is a useful measure for the extendedness of the chain and accessible experimentally by means of the fit.

At this point already – to be discussed in further detail later – it is obvious that the contact probabilities are strongly increased by the existence of the chromophores even though the random coil distributions are similar between the X_{16} and X_{14}ct systems. This increase was to be expected due to the attraction between the chromophores being about 2.5-fold stronger than the Q–Q and S–S interactions, but now it can also be quantified.

The four parameters b, σ, P_{cnt} and $\sqrt{\langle d^2 \rangle}$ all depend on the temperature. This temperature dependence is the key to finding ϵ_{HB}: the experimental values were measured at room temperature, so the reduced temperature at which all four parameters equal the experimental results would be called 300 K. Unfortunately, an exact agreement of all parameters is never reached, so the procedure is not that simple. Instead, it becomes necessary to look at the temperature dependences themselves, shown in fig. 6.2. The graphs show b (continuous lines) and σ (dashed) for the four systems from fig. 6.1 as well as Q_{14}c and S_{14}c. $\sqrt{\langle d^2 \rangle}$ has been left out to reduce visual noise and P_{cnt} is shown in fig. 6.4 to be discussed in more detail there.

Again, the variation in b and σ (and $\sqrt{\langle d^2 \rangle}$) between the three chain modifications is small, in contrast to P_{cnt}, where the graphs differ wildly. Both b and σ are vaguely stable at high T. Nearing the transition temperature, shown as a vertical dotted line for each system, b drops to very low values (equivalent to the maximum of $p_T(d)$ moving to lower d) and σ tends upwards (equivalent to the distribution becoming wider, counterbalancing the effect of b), although for S_{14}c and S_{14}ct it even decreases again in the vicinity of T^*.

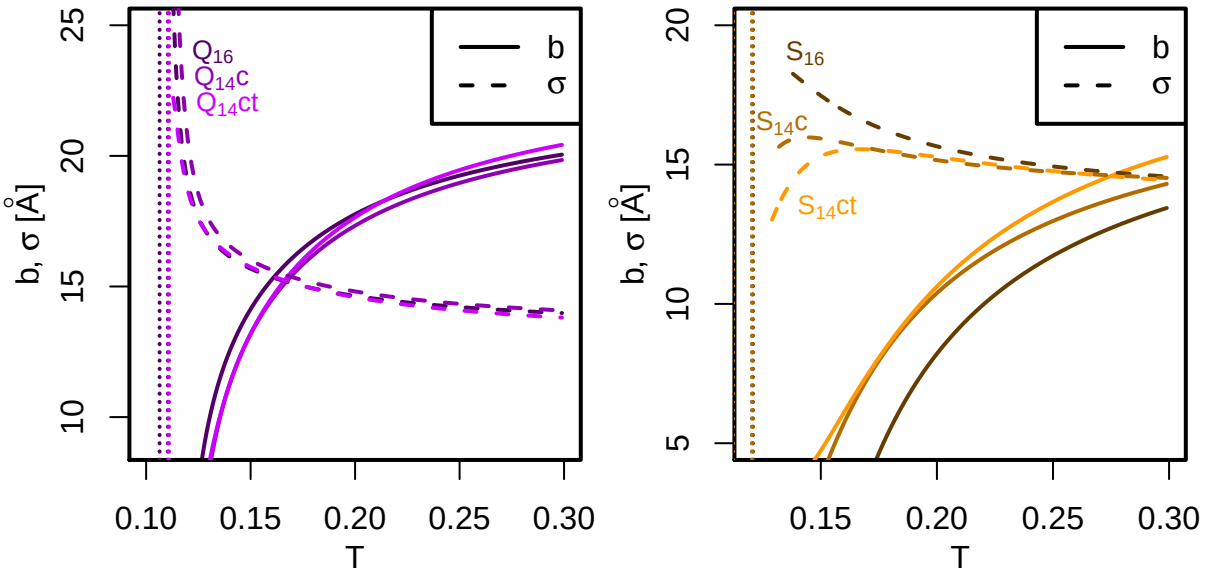

Figure 6.2 – Temperature dependence of the fit parameters b and σ for Q_{14}- and S_{14}-based systems.

The experimental values $P_{\text{cnt}} = 0.15$ and $\sqrt{\langle d^2 \rangle} = 25.9$ Å [62] as well as b and σ [P. Enke, private communication[1]] establish a relatively broad interval of temperatures because Q_{14}ct reaches a contact probability of 15% only at $T = 0.115$, but a chromophore distance as large as 25.9 Å is not observed below $T = 0.16$. The b and σ graphs begin to stabilise around the latter temperature; the former is very close to the collapse transition. Within this interval, the needed assignment of $T' = 300$ K to a fixed T can only be estimated subjectively.

Walters et al. [214] found a lower chromophore distance around 22.5 Å in a similar set-up, which is an argument in favour of lower $\sqrt{\langle d^2 \rangle}$ and therefore motivates a decision towards the lower end of the interval. Similarly, the PRIME20n analysis suggested that PRIME20 tends to underestimate $\sqrt{\langle d^2 \rangle}$. On the other hand, both polyQ and polyS are reported to be mostly disordered in a variety of publications, suggesting that the usual experimental situation is not too close to transition conditions. Thus, the collapse transition should not lie too close to 300 K. In the end – taking the uncited b and σ into account as well – $T' = 300$ K was chosen to be mapped to $T = 0.135$ for Q_{14}ct and for all further polyglutamine sequences. Hence, T'/T equals 2222 K and $\epsilon_{\text{HB}} = 4.4$ kcal/mol.

Following similar arguments for S_{14}ct, the physical temperature of 300 K was mapped to $T = 0.160$ for polyserine[2], so $T'/T = 1875$ K and $\epsilon_{\text{HB}} = 3.7$ kcal/mol.

Both of these ϵ_{HB}values are much lower than that of polyA found before. At first glance, they seem reasonable in comparison to the literature values clustering around 5 kcal/mol, but if the temperature shift in PRIME20n is taken into account, they become approximately 2.9 kcal/mol and 2.4 kcal/mol instead, which is unexpectedly low. The lack of hydrogen bonding capability between side chain and backbone in the model may be the cause of these values. Regarding ϵ_{HB} as a product of the competition for hydrogen bonds between polar groups in the backbone, side chain and solvent, this interaction is a factor which the model neglects.

Notwithstanding this physical issue, the T' scales achieved by this analysis match the experimental templates reasonably well and will be used for the remainder of the chapter.

[1] The results for b and σ are yet to be published at the time of writing.

[2] Based on unpublished results by P. Enke as well. Earlier publications from the same lab on polyserine [103, 104] do not name the values explicitly.

6.2 Transition temperatures

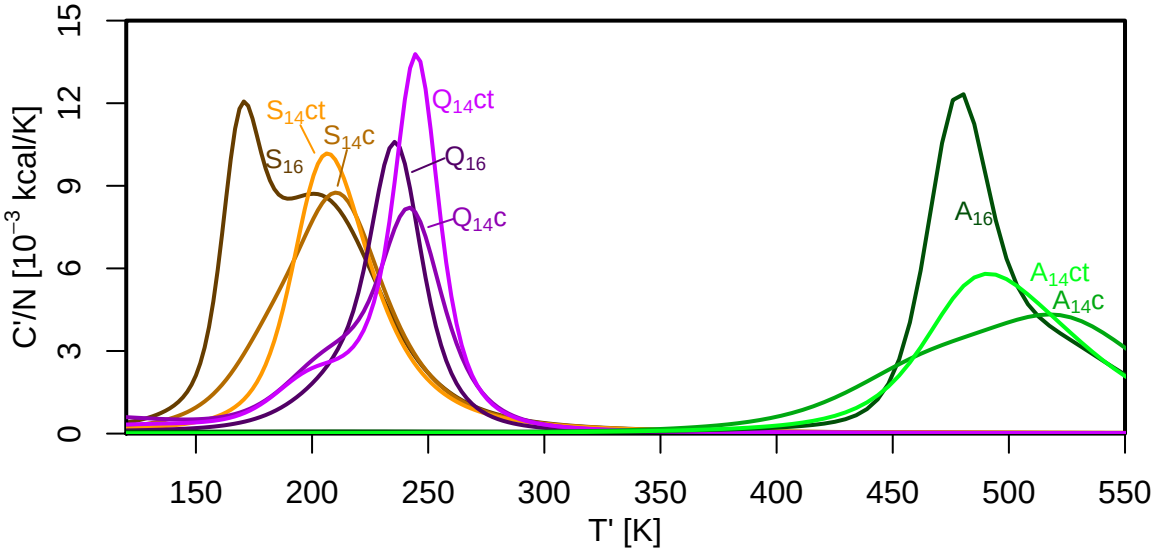

Figure 6.3 – Heat capacities of peptides with and without chromophores and tail versus physical temperature.

Having established an individual temperature scale for each type of peptide, it is now possible to study their thermodynamics and structure formation depending on the modification by chromophores and the solubility tail. Once again, the canonical heat capacity is a useful overview of the thermodynamic properties of the systems and it is depicted in fig. 6.3 for all nine systems, i.e. three modifications of polyA, polyQ and polyS respectively. As in chapter 4, the microcanonical analysis yields more detailed results, but such detail is not of use here, so the existence of additional states should be acknowledged, but not regarded further. (The microcanonical heat capacities are depicted in the appendix in fig. B.3.)

The maxima of all three polyalanine $C'(T')$ graphs are shifted far to the right in comparison to those of polyglutamine and polyserine. This is of course due to the much higher ϵ_{HB} value and not unexpected, polyA is in its helical state at 300 K while polyQ and polyS are disordered at this temperature.

Within each family of three polyX modifications (X=A,Q,S), the $C'(T')$ maxima follow a similar pattern: the maximum of $X_{14}c$, the peptide with added chromophores, lies at a higher position than that of X_{16} and its peak height is diminished. Upon adding the solubility tail ($X_{14}ct$), the temperature remains approximately stable, defying the intuition that an increase in chain length should increase the collapse temperature as well. In the case of polyA, the temperature even decreases again. This behaviour indicates that the tail indeed increases solubility, even in this implicit solvent simulation. The peak height of $X_{14}ct$ is increased in comparison to $X_{14}c$.

Table 6.1 shows the minimum energies U_0 and U_m for all chains considered here (including $N = 9$), as already seen in tables 4.2 and 5.3. What catches the eye is that in both helical cases (polyA and polyS) the lowest converged energies U_m of $X_{14}c$ are consistently higher than those of X_{16}. This explains the diminished $C'(T')$ peak height because $C'(T')$ is the derivative of the internal energy $U_c'(T')$, hence a smaller difference in U_c' generally produces smaller values in C'. When the solubility tail is added, the U_m values[3] are lower again due to the increased number of residues. The latter effect applies to U_0 as well. Notably, the U_0 values of $X_{14}c$ and X_{16} are mostly similar despite the U_m discrepancies, indicating a slower convergence behaviour.

[3] U_c is a canonical quantity while U_m and U_0 are microcanonical, but their value ranges and physical significance are identical, so they can be used interchangeably in this argument.

Table 6.1 – Lowest observed potential energies U_0 and lowest potential energies U_m at which a converged $g(U)$ could be acquired of three modifications of $N = 9$ and $N = 14$ peptides. U_0 are values obtained from single configurations; U_m lower boundaries of the energy bins used in simulation. All values are given in reduced units to comply with the format of tables 4.2 and 5.3.

	polyA		polyQ		polyS	
Modification	U_0	U_m	U_0	U_m	U_0	U_m
X_{11}	-8.092	-7.6	-7.960	-5.3	-8.462	-8.3
X_9c	-8.533	-5.0	-6.791	-5.7	-8.925	-5.0
X_9ct	-12.112	-10.0	-11.203	-9.5	-13.043	-10.0
X_{16}	-13.512	-12.2	-9.160	-7.4	-14.182	-14.0
$X_{14}c$	-13.916	-10.0	-10.786	-9.6	-13.838	-10.0
$X_{14}ct$	-16.152	-12.0	-15.500	-13.4	-17.763	-12.0

Aside of the influence on $C'(T')$, the observation of narrower energy ranges itself is an interesting hint at the structure formation. As established, polyA and polyS form γ-helices as their native states and these helices are the configurations of lowest possible energy. Upon addition of chromophores, these helices are most likely still available, as the low U_0 values suggest, but a worsened convergence behaviour of the SAMC simulation indicates that the tendency to form such helices is lower than in the unmodified systems. The resulting structures will be discussed in section 6.4.

6.3 The random coil state

In advance of a discussion of folded states, the random coil state should receive some further attention because this is the state in which the FRET/TTET experiments of polyQ and polyS take place.

As the heat capacities (fig. 6.3) indicate, all of the investigated polyQ and polyS systems collapse at temperatures far below 0 °C because the temperatures have been scaled such that the distance distributions $p_T(d)$ in fig. 6.1 best reproduce the experimental behaviour. This experimental behaviour was interpreted to be coil-like and sensible to be fitted with the Edwards model, hence the collapse must occur at lower T'. The calculated contact probabilities of about 15% support this interpretation because in a folded, ordered state, they should either tend towards zero or towards one. Both of these values are imaginable – for example, a helical configuration would lead to a contact probability of zero, a hairpin to a probability of one. But a value like 15% indicates the absence of either structure, hence a disordered state. Furthermore, polyQ is generally considered as an intrinsically disordered protein and polyS is vaguely described to be disordered in literature as well, either a random coil or a denser globule, but not regularly folded at room temperature.

In this unfolded state, it is interesting to note that the contact probabilities (see fig. 6.1) are still much lower than 15% (traded for realistic b and σ parameters in the T scaling procedure, as discussed along fig. 6.2) and they differ strongly between the systems. The strong attraction between the chromophore side chains is expected to increase the contact probability. In $S_{14}c$ and $Q_{14}c$ at 300 K, these probabilities are 0.073 and 0.069, much higher than the 0.029 and 0.014 of S_{16} and Q_{16}. The solubility tail increases conformational freedom of the chain as a whole and should reduce P_{cnt} again. This effect is observed as well, with probabilities of 0.066 and 0.059 in $S_{14}ct$ and $Q_{14}ct$ at this temperature. It is noteworthy that these values are closer to those of $S_{14}c$ and $Q_{14}c$ than to S_{16} and Q_{16}, supporting the initial hypothesis that the tail should not affect the configurations too much (at least in the random coil state).

Despite the differences in contact probability, the out-of-contact regions of the graphs do not differ too much and all systems are described equally well by the random coil picture. It can be argued that

most further experimental behaviour, especially regarding dynamics, is not too strongly influenced by these modifications to the basic systems in question. Further confirmation, using a better model and perhaps also adding a comparison of dynamic properties, would be desirable, but cannot be achieved in the framework of the present project.

6.4 Folded states

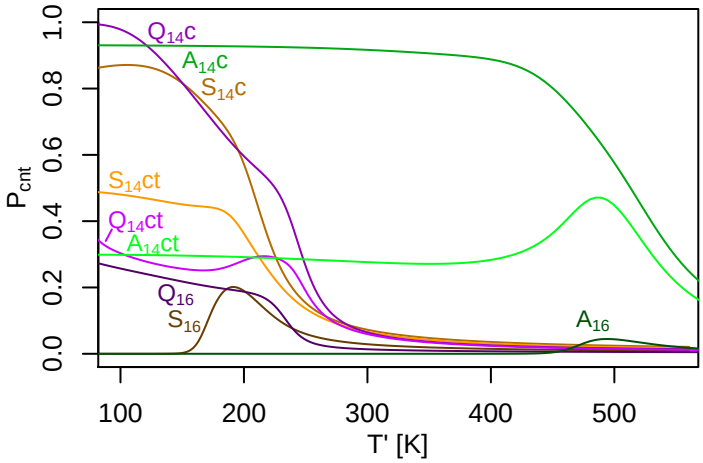

Figure 6.4 – Terminal side chain contact probabilities of peptides with and without chromophores and tail versus physical temperature.

Unlike the random coil states, the folded states experience considerable deformations by the chromophores and solubility tail. Figure 6.4 shows the chromophore contact probabilities $P_{\text{cnt}}(T')$ of the nine systems (cf. eq. (3.30)). Among these, the unmodified cases have already been discussed in chapter 4. The contact probability of polyA and polyS is zero in their low-T' helix conformations while Q_{16} reaches contact probabilities around 30%. This value may seem surprisingly low for a hairpin state, but as seen earlier, this state allows a certain level of flexibility and the terminal side chains will not necessarily interact in such configurations.

Upon adding the chromophores, the low-T' contact probabilities increase to between 80% and 100%. Evidently, these folded configurations are dominated by the chromophore contact, which is slightly surprising because the energy gain of a contact between the chromophores is far less than that of an H-Bond. For polyQ, it is well conceivable that the hairpin configurations which include this contact are now preferred versus those without it; for polyS and polyA, the helix state should still be energetically favourable, but as discussed earlier, the simulations did not converge to these energies. The helix-globule folding transition, originally closely below the coil-globule collapse, seems to have been shifted to a much lower temperature by addition of the chromophores.

Finally, P_{cnt} of the X_{14}ct systems with chromophores and tail follows the X_{14}c graphs at high temperatures, but at the collapse/folding transition begins to deviate and stabilises between 30% and 50% at low T'. These values are not far from the original probabilities in polyQ, but an extreme change in polyA and polyS, indicating a strongly reduced propensity towards helix formation, at least at these temperatures.

The low-T' configurations (at 100 K) of all nine systems are found as HB matrices in fig. 6.5 and as snapshots in fig. 6.6. The top row in both figures depicts polyQ, the middle row polyS and the bottom row polyA, and the modifications are X_{16}, X_{14}c and X_{14}ct from left to right in each row. In the X_{14}ct matrices, two lines mark the solubility tail: the 16x16 square comprising the majority of these matrices depicts the main chain and lends itself for comparison to the X14c and X_{16} matrices. The bottom right

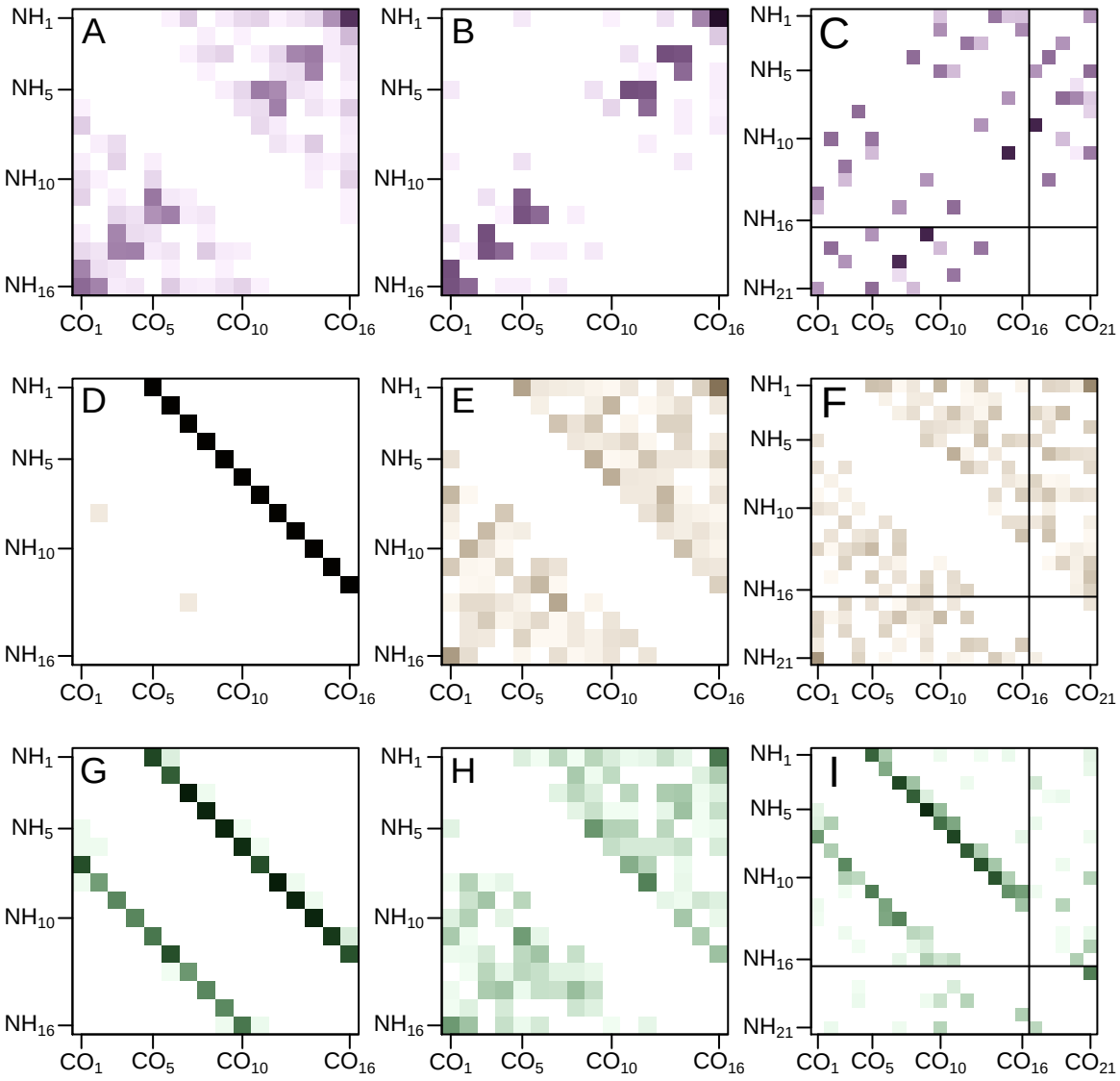

Figure 6.5 – H-Bond contact matrices of peptides with and without chromophores and tail at $T' = 100\,\mathrm{K}$. The systems are (A) Q_{16}, (B) $Q_{14}c$, (C) $Q_{14}ct$, (D-F) polyS and (G-I) polyA in the same order. In subfigures C, F and I the solubility tail is separated from the main chain by black lines.

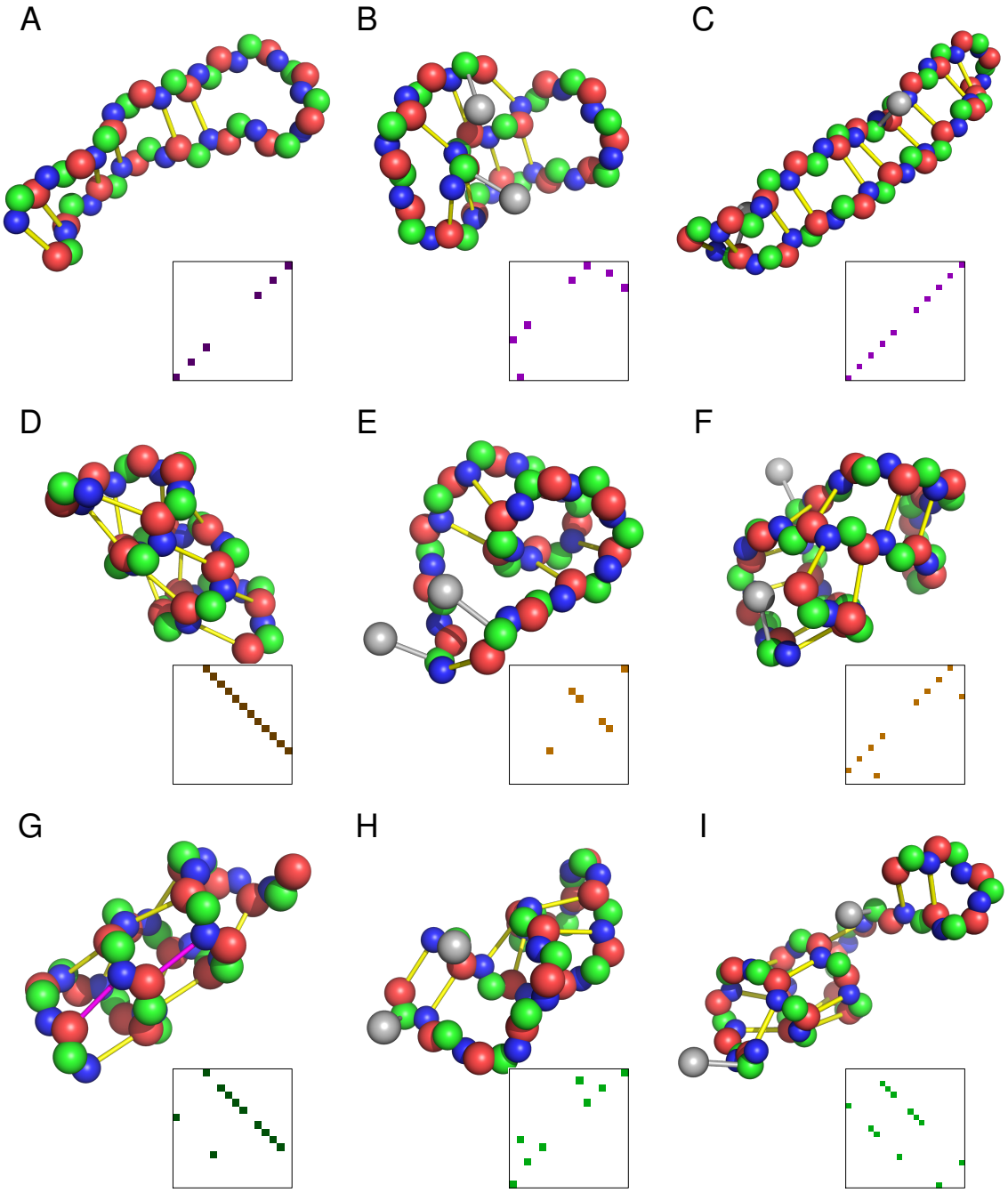

Figure 6.6 – Configuration snapshots and corresponding HB matrices, arranged in the same way as fig. 6.5. Chromophore side chains are shown as grey spheres.

5x5 square presents interactions within the tail (of which only the $NH_{17} \cdots CO_{21}$ and $NH_{21} \cdots CO_{17}$ contacts are allowed in PRIME20) and the two remaining rectangles include interactions between tail and main chain residues. In the snapshots, the tryptophan (chromophore) side chains are shown as grey beads and sticks; all other side chains are left out for ease of view.

Beginning at the polyglutamine HB matrices (top row), the visual difference between Q_{16} and $Q_{14}c$ is surprisingly small given the large deviation of P_{cnt} between the two systems. Both systems fold into the documented hairpin state; in $Q_{14}c$ its contrast is enhanced in comparison to Q_{16}, indicating a larger fraction of hairpins as well as a stronger preference for the main and first side diagonals. At this temperature of 100 K, more than 90% of the $Q_{14}c$ configurations are hairpins and less than 5% lie on the more distant side diagonals, while for Q_{16} these numbers are about 70% and 13%. In the configurations on the main and first side diagonals, named "Q_{16}-" and "Q_{15}-based" hairpins in chapter 4, the distance between the chain termini is small, enabling the chromophore contact. In the Q_{14}- or Q_{13}-based configurations on the distant side diagonals, the chain ends do not meet and the desirable chromophore contact is not formed, which leads to the observed configuration. The $Q_{14}c$ snapshot in fig. 6.6 shows another curious configuration, a slightly deformed double-hairpin variant, which also allows the chromophore contact to be formed.

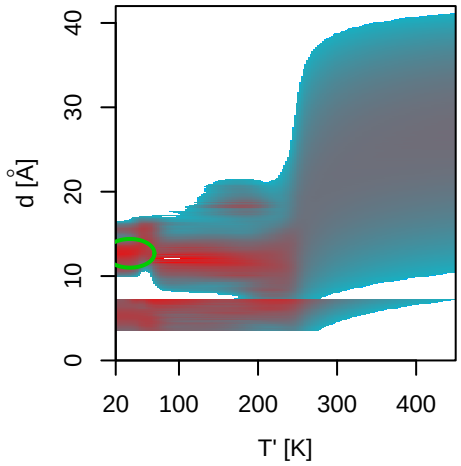

Figure 6.7 – Chromophore distance distribution $p(d, T')$ of $Q_{14}ct$. The green ellipse highlights a low-T' feature at 12.7 Å.

The $Q_{14}ct$ matrix and snapshot (top right in both figures) highlight one of the hazards of the asymmetric choice of a solubility tail compared to symmetric flanking residues at both termini. Although the main chain section of the matrix still contains a hairpin signature, it does not dominate the matrix as much as it does in the other two modifications. In addition, it appears to be part of a double-hairpin signature, with the tail participating in the configuration as well. Another interesting population in this figure spans the main diagonal without any discrimination between main chain and tail. It indicates a long hairpin which incorporates both sections of the peptide, as it is shown in the snapshot. The chromophores are obviously not in contact – in this case the increased number of H-Bonds outweighs their importance. Instead, they are found at a characteristic distance of 12.7 Å, which is prominently visible in the $p(d, T')$ graph as well (highlighted in fig. 6.7) and has been observed similarly in MD simulations of the same system [166]. $p(d, T')$ also clearly indicates a coexistence at low temperatures of configurations like this and hairpin or double-hairpin configurations in which the chromophores are in contact.

Regarding the helical peptides, the relation between the $(A/S)_{16}$ and $(A/S)_{14}c$ is the same in both cases. As already speculated in the discussion of P_{cnt} (fig. 6.4), the chromophore interaction overrules helix formation despite its smaller energetic contribution. The matrices of A_{16} and S_{16} in the left column exhibit the documented descending $NH_i \cdots CO_{i+4}$ diagonal characteristic for γ-helices. In the $(A/S)_{14}c$ matrices, this signature is not visible at all. The chromophore contact is prominent, and a few further contacts stand out, but no regularity is seen. The corresponding snapshots are therefore not to be regarded as typical structures; they only present two out of many examples how the systems are able to form configurations incorporating the chromophore contact: a single helix turn of $S_{14}c$ with the ends stretching out to form the contact, and a twisted hairpin of $A_{14}c$, curiously with eight H-Bonds instead of the usual six.

Similarly to $Q_{14}ct$, the $S_{14}ct$ H-Bonds do not discriminate at all between main chain and tail residues. This is not surprising as the tail contains two serine residues itself and, as seen in table 2.3 in the model

chapter, the interaction strength between the serine and arginine side chains is equal to that between two serines. Hence, the only distinction between main chain and tail is the larger size of the two arginine side chains in comparison to those of serine. In the HB matrix, this leads to the same disorder as in S_{14}c. The system could be imagined to form a γ-helix incorporating both main chain and tail (similar to the long Q_{14}ct hairpin), but such a structure is not prominent in the HB matrix, either due to a failure in reaching low energies during simulation or due to a repulsion between the large and charged arginine side chains disrupting the helix structure.

In A_{14}ct, the picture is different. The hydrophobic alanine side chain has a repulsive interaction with the polar serine and arginine side chains, causing the tail to act separately from the main chain. Therefore, the main chain is able to form its usual γ-helix (again indicating that this helix should be achievable by A_{14}c as well) and the tail bends back onto itself, closing the only two H-Bonds possible within the limits of PRIME20. This bipartite structure is seen in the snapshot in the bottom right corner of fig. 6.6 as well. The chromophore interaction does not play a role in this case, which underlines that this positioning of chromophores is unsuitable for the investigation of helical peptides.

In conclusion of this chapter, the chromophores and solubility tail evidently have a strong influence on the folded states of these peptides. This influence depends on the properties of the peptide in question. More specifically speaking, the polar tail can interact favourably with the main chains of polyS and polyQ and is therefore incorporated in structure formation, while in the case of polyA the interaction is unfavourable and the tail tends to separate from the main chain. The attraction between the chromophores tends to destabilise helix conformations of polyA and polyS and to stabilise the polyQ hairpins. If helix formation is expected in experiment (like for polyA), different chromophore positions like an $(i, i + 6)$ spacing are more suitable [63, 138].

At higher temperatures, in the random coil state, the influence is smaller. The tail does not appear to affect the global properties of the peptides much, confirming the initial argument. The chromophores have a higher probability to be in contact than non-chromophore residues in the same positions. This effect needs to be considered for the interpretation of experimental results, but the out-of-contact structures are largely unaffected by the chromophores. Since both polyQ and polyS are unfolded at room temperature, the influence on folded states does not play an important role. These observations largely justify the use of such chromophores and of the solubility tail in the loop formation experiments, but possible results at low temperatures, near the folded states, need to be treated with care.

Chapter 7

Conclusion

The preceding three chapters presented three facets of the problem of protein folding. The behaviour of the example systems polyglutamine, polyserine and polyalanine was investigated depending on chain length, choice of model and choice of experimental factors using the intermediate-resolution protein model PRIME20 for a balance between computational simplicity and molecular detail and the algorithm SAMC, which promises complete thermodynamic information of the simulated system.

In the first part (chapter 4), both polyalanine and polyserine were found to fold into a helical state at low temperatures. Regarding polyserine, which is usually considered to be disordered or even helix-breaking, the structure came as a surprise, but during the assignment of physical temperatures to the reduced temperature scale in chapter 6, the polyserine folding temperature was interpreted to lie below the freezing point of water, which is why the finding does not contradict those observations of disorder.

The polyalanine folding temperature on the other hand lies far higher, around 462 K. In simulation, the transition temperatures of both systems are similar, hence the conversion factor between the temperature scales must vary strongly. This factor relates to the effective strength of backbone hydrogen bonds ϵ_{HB}, which evidently is affected by the ability of side chains to form hydrogen bonds with the backbone or solvent because these interactions are not expressed explicitly by the model. Reasonable ϵ_{HB} values for all three systems were obtained in chapter 6, but these values are not general in any respect, so simulations of other polypeptides will require new individual ϵ_{HB} again.

In the model chapter on page 10, a word of warning by Miyazawa and Jernigan was quoted in which they note that knowledge-based models like PRIME20 are inherently biased towards collapsed configurations. The helix structure of polyserine can also be interpreted as a consequence of this bias, counterbalanced by the unusually low ϵ_{HB} value, which leaves the molecule in its random coil state at room temperature.

A second surprise was the identification of the folded states of polyalanine and polyserine to be γ-helices. The γ-helix structure had been postulated concurrently with the α-helix, but has never been observed in a real protein since then. The finding of stable γ-helices here, accompanied by the complete absence of α-helices, made the results highly questionable and required a much more thorough analysis of the model parameters than initially expected.

This analysis was performed in chapter 5. The previous formulation of PRIME20 turned out to be incomplete, lacking accurate so-called squeeze parameters. These parameters describe modifications to the effective diameters of beads if they are separated by a short distance along the chain. Two classes of squeeze parameters were treated separately, one modifying only backbone beads, the other modifying side chain beads in interaction with the backbone.

Most of the backbone squeeze parameters are factors smaller than one, so the effective bead sizes are reduced. The incomplete version of the model already contained an approximate variant of some of these factors, which is smaller than the accurate values. Thus, some configurations of the backbone –

including the α-helix – were found to be accessible in the new version of PRIME20 (dubbed PRIME20n), but other configurations were found to be inaccessible now. This result from a theoretical treatment of Ramachandran plots was confirmed by the observation of α-helices in simulations of polyalanine and polyserine using the new model variant.

The second class of squeeze parameters governs interactions between side chain and backbone beads. These factors mostly lie between 1 and 1.5, corresponding to a substantial increase of the effective bead diameters. In the Ramachandran plot analysis, they turned out to restrict the available configuration space of a single residue (or of a random coil) to just a fraction of the space in PRIME20 and most notably to overlay the γ-helical regions of the plot. Supporting this analysis, such helices were no longer observed in the simulations of polyserine and polyalanine applying PRIME20n.

A hypothetical third model, using the side chain squeeze parameters, but not those of the backbone, was studied as well. According to the analysis of Ramachandran plots, it was expected not to form any helices, and indeed neither of the α and γ types occurred. A single simulation found a π-helix instead, and the remaining configurations were on hairpins or looser turn structures.

Polyglutamine mostly remained in hairpin configurations throughout all models and chain lengths. In a sequence of more than 8 amino acids, these configurations have a higher potential energy than α- or γ-helices, but both helix types are mostly inaccessible for glutamine residues due to the larger side chains compared to alanine or serine. A few simulation runs reached γ-helical polyglutamine configurations in PRIME20 and α-helix fragments in PRIME20n, but due to the strong geometric restriction (often called a "bottleneck" of simulation), the SAMC algorithm did not converge to the corresponding energies. It was therefore impossible to assign accurate temperatures to these potential states. Like polyserine, polyglutamine is intrinsically disordered at room temperature for the considered chain lengths (but fig. 8.1 indicates a change in behaviour for longer chains), so the folding transition temperatures are below 0°C already and the helix transitions would lie much lower. Thus, these structures are only relevant from a modelling perspective, but not for a biological interpretation.

In chapter 6, the influence of dyes needed for spectroscopy experiments and of a tail which is essential to keep peptides like polyglutamine soluble in aqueous solvents was studied. To this end, chains with and without these dyes (attached to both termini) and tail (at the C-terminus) were simulated. The dyes, large synthetic amino acids (or molecules attached to the chain through amino acid linkers), were represented by tryptophan, the largest proteinogenic amino acid. The tail, a sequence of polar proteinogenic amino acids, could be expressed faithfully in PRIME20.

For polyglutamine and polyserine, a strong attraction between the dyes as well as favourable interactions between the main chain and the tail were found to cause significant disruptions of the respective folded states, effectively destroying the helical state of polyserine and adding further hairpin varieties to polyglutamine. For polyalanine, although the dyes hampered helix formation, the observed effect of the tail was much smaller because the polar residues avoided interaction with the hydrophobic main chain.

In the random coil state on the other hand, none of the systems were strongly affected by the additions. Although the probability to form a contact between the N- and C-terminal side chains was noticeably increased by the dyes, the out-of-contact distance distribution was mostly similar between the systems with and without dyes and mostly unaffected by the tail as well. Thus, the described spectroscopy experiments can be considered reliable for disordered peptides like polyglutamine, but folded states have to be handled with care.

Chapter 8

Outlook

8.1 A modified PRIME20

As explained before, the third results chapter is based on results obtained with the "old" PRIME20 force field rather than PRIME20n because this model variant was not known at the time of these simulations. Even though the observed effects were argued not to depend too strongly on the force field, a sense of uncertainty remains, and given enough time it might have made sense to redo the simulations with PRIME20n. However, while the results regarding Ramachandran plots and helix formation are undoubtedly closer to reality in PRIME20n than in PRIME20, the model has a number of weaknesses which need to be addressed before further Monte Carlo simulations can be performed.

One weakness pertains to the existence and values of the aforementioned squeeze parameters. As seen, if the unsqueezed bead diameters were used, an overlap between backbone CO beads would render all configurations illegal. On the other hand, reducing all bead sizes results in an unrealistic amount of flexibility, as shown by Rutter's analysis of the model [174]. The squeeze factors might therefore prove to be a necessity, but the values assigned to them seem arbitrary, especially in comparison to the energy parameters. While these are the results of an effortful optimisation process based on PDB structures, no explanation regarding the origins of the squeeze parameters can be found in published literature (and in fact finding the factors themselves in literature is only possible by virtue of an accurate knowledge of what one seeks), and for example the value of 0.87829 for $sqz3$ yields an effective bead diameter with an accuracy equal to $1/1\,000\,000$ of the Bohr radius. Technically this number is of course irrelevant, but it certainly does not match the philosophy of a moderately coarse-grained model which PRIME20n claims to be, and the arbitrary, seemingly on-the-fly squeeze factors present a stark contrast to the well-documented energy values. Worse still, while squeeze factors smaller than one have been argued convincingly to express the united-atom property of the backbone beads, $sqz1$ is larger than one and is not explained by this argument at all. The same issue applies to most of the side chain squeeze factors as well.

This problem can be argued to be a purely philosophical one, as the technical process of simulating the model is not affected drastically by the squeeze factors. A bigger problem, especially for Monte Carlo simulation, arises through the auxiliary interactions which are used to stabilise hydrogen bonds. The Monte Carlo procedure consists of randomly selecting a sterically legal configuration, then calculating its potential energy and finally accepting or rejecting the configuration based on this energy. A prerequisite for such a procedure is a model in which every configuration is unambiguously assigned one potential energy. The auxiliary interactions violate this condition: a configuration in which an H-Bond could be formed, but is hindered by these interactions, can either be considered legal without the H-Bond, or illegal with the H-Bond, leading to different energies as well as a different selection of H-Bond partners available for further interaction. In a molecular dynamics simulation, such a configuration is interpreted

depending on its history, i.e. on the question whether the H-Bond was already formed in the previous time step. In Monte Carlo, the concept of a "previous time step" does not exist and hence the treatment of such a configuration remains ambiguous, hindering ergodic sampling of configuration space.

On a similar note, an eleventh squeeze parameter reduces the bead distance of NH and CO beads involved in an H-Bond. Like the auxiliary interactions, this squeeze parameter is a history-dependent size change and is neither compatible with a Monte Carlo approach, nor with the underlying concept of ergodicity.

Finally, the model description explicitly states that an H-Bond can only be formed by NH and CO beads with three intervening residues. As noted in the first results chapter, this restriction clashes with the definition of a β-turn, which contains an $\mathrm{NH}_i \cdots \mathrm{CO}_{i-3}$ H-Bond [152, 200]. Although the Ramachandran plots in the second results chapter (fig. 5.5 and similar) indicate that the geometry already prevents the so-called β_{III} turns (synonymous with 3_{10} helices), the non-repeating $\beta_{I/II}$ types could not be considered in that analysis and might still be available geometrically. Regardless of this Ramachandran plot analysis, these turns are a crucial element of secondary structure and should be made available geometrically as well as energetically.

Considering these issues, a further use of PRIME20n (or PRIME20) does not appear sensible. However, discarding the model entirely would not do the effort justice which has been put in its creation and especially led to the still elegant set of energy parameters. Therefore, an attempt to produce a variant of PRIME20 has already been initiated, in which the above issues are supposed to be resolved without impairing the quality of resulting thermodynamics and structures.

8.2 Longer peptides and aggregation

This dissertation concludes the funding period of a research project whose original grant proposal neither contained the investigations of PRIME20n nor the influence of chromophores. Both of these aspects came up during the project time. Instead, the original aim was to study the structure formation of the "Alzheimer peptide" Amyloid-β (Aβ) and its oligomerisation process. This aim is still relevant and therefore a possible future topic of simulation.

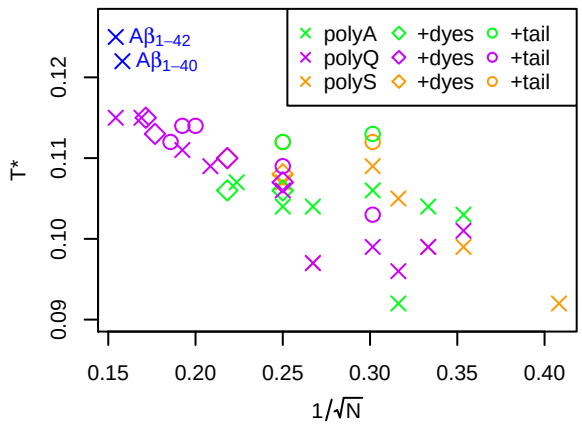

Figure 8.1 – Chain length dependence of the transition temperature T^* of all systems simulated, including preliminary simulations of long polyglutamine chains and Amyloid-β.

Regarding the single-chain behaviour of Aβ, it may also be instructive to compare it to further peptides of similar length. In the previous results chapters, only peptides up to length $N = 23$ were considered. Preliminary simulations of longer polyglutamine chains ($N = \{25, 27, 35, 42\}$) and of two Aβ variants ($N = \{40, 42\}$) were performed in the past and discontinued because of the changed research focus and because they had employed the outdated PRIME20 model. Figure 8.1 shows the transition temperatures of all these systems (in PRIME20) versus chain length. The figure contains the polyA, polyS and polyQ chains regarded in earlier chapters, the longer polyglutamine and Aβ peptides and the experimental modifications of three medium-length polyglutamines, i.e. chains with added spectroscopic dyes and the solubility tail. The horizontal axis is scaled in accordance with the mean-field expectation of a $T^* \propto N^{-1/2}$ dependence [169], and a visual inspection confirms the existence such a dependence for the longer chains, seen by a roughly constant

negative slope in the graph. Also, due to the transition temperatures increasing with N, it is conceivable that the longer polyglutamines may be collapsed at physiological temperatures. Such a result would have relevant implications for the polyQ diseases, for example because the folded structures may act as starting points of aggregation, as has been discussed elsewhere. A more rigorous analysis of these preliminary data from an outdated model would be useless, so the visual trend shall suffice for this outlook. The Aβ transition temperatures are clearly higher than the general trend, which may be an indication of a different behaviour, but may also be caused by insufficient energy ranges in these low-effort simulations. Such questions can be addressed in a continuation of the project.

Finally, the program code can be expanded to investigate aggregation of multiple peptides. Here, of course Aβ and polyglutamine are of interest in order to better understand the polyQ disease threshold, the structures of Aβ and their influence on Alzheimer's disease. Additionally, the thermodynamic stability of fibrils or oligomers is an interesting topic for future research: while the pathological amyloid plaques of Aβ or polyQ appear to be inert, the same structures can be functional. The parathyroid hormone (PTH) is stored in an amyloid form, but can be reverted into its native state by a change of conditions which has not been understood yet. A complete picture of the thermodynamics and structures of single PTH chains or oligomers in comparison to Aβ or polyQ might pave a way towards understanding the storage and access procedure and perhaps even towards a treatment of amyloid diseases.

Bibliography

[1] Worldwide Protein Data Bank. Url: www.wwpdb.org, last visited 2019/01/27.

[2] IONIS-HTT Rx (RG6042) top-line data demonstrate significant reductions of disease-causing mutant Huntingtin protein in people with Huntington's Disease. Press release, 2018. Url: http://ir.ionispharma.com/node/23401/pdf, last visited 2019/01/31.

[3] A. Abu-Baker and G. A. Rouleau. Polyalanine and polyglutamine diseases: possible common mechanisms? In R. D. Wells and T. Ashizawa, editors, *Genetic Instabilities and Neurological Diseases*, chapter 41, page 645. Academic Press, second edition, 2006.

[4] A. Adegbuyiro, F. Sedighi, A. W. Pilkington, S. Groover, and J. Legleiter. Proteins containing expanded polyglutamine tracts and neurodegenerative disease. *Biochemistry*, **56**:1199, 2017.

[5] A. Albrecht and S. Mundlos. The other trinucleotide repeat: polyalanine expansion disorders. *Curr. Opin. Genet. Dev.*, **15**:285, 2005.

[6] E. L. Altschuler, N. V. Hud, J. A. Mazrimas, and B. Rupp. Random coil conformation for extended polyglutamine stretches in aqueous soluble monomeric peptides. *J. Pept. Res.*, **50**:73, 1997.

[7] C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, **181**:223, 1973.

[8] A. M. Anton, A. Heidebrecht, N. Mahmood, M. Beiner, T. Scheibel, and F. Kremer. Foundation of the outstanding toughness in biomimetic and natural spider silk. *Biomacromolecules*, **18**:3954, 2017.

[9] A. M. Anton, W. Kossack, C. Gutsche, R. Figuli (Ene), P. Papadopoulos, J. Ebad-Allah, C. Kuntscher, and F. Kremer. Pressure-dependent FTIR-spectroscopy on the counterbalance between external and internal constraints in spider silk of Nephila pilipes. *Macromolecules*, **46**:4919, 2013.

[10] E. A. Asante, M. Smidak, A. Grimshaw, R. Houghton, A. Tomlinson, A. Jeelani, T. Jakubcova, S. Hamdan, A. Richard-Londt, J. M. Linehan, S. Brandner, M. Alpers, J. Whitfield, S. Mead, J. D. F. Wadsworth, and J. Collinge. A naturally occurring variant of the human prion protein completely prevents prion disease. *Nature*, **5224**:478, 2015.

[11] W. T. Astbury, S. Dickinson, and K. Bailey. The X-ray interpretation of denaturation and the structure of the seed globulins. *Biochem. J.*, **29**:2351, 1935.

[12] W. T. Astbury, H. J. Woods, and W. L. Bragg. X-Ray studies of the structure of hair, wool, and related fibres. II.- the molecular structure and elastic properties of hair keratin. *Philos. Trans. Royal Soc.*, **A232**:333, 1933.

[13] R. Aurora, T. P. Creamer, R. Srinivasan, and G. D. Rose. Local interactions in protein folding: lessons from the α-helix. *J. Biol. Chem.*, **272**:1413, 1997.

[14] O. T. Avery, C. M. MacLeod, and M. McCarty. Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Inductions of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *J. Exp. Med.*, **79**:137, 1944.

[15] R. L. Baldwin. Matching speed and stability. *Nature*, **369**:183, 1994.

[16] R. L. Baldwin. The search for folding intermediates and the mechanism of protein folding. *Annu. Rev. Biophys.*, **37**:1, 2008.

[17] J. Baschnagel, J. P. Wittmer, and H. Meyer. Monte Carlo simulation of polymers: Coarse-grained models. In N. Attig, K. Binder, H. Grubmüller, and K. Kremer, editors, *Computational Soft Matter: From Synthetic Polymers to Proteins*, chapter 4, page 83. John von Neumann Institute for Computing, 2004.

[18] P. A. Bates and N. A. DeLuca. The polyserine tract of herpes simplex virus ICP4 is required for normal viral gene expression and growth in murine trigeminal ganglia. *J. Virol.*, **72**:7115, 1998.

[19] T. Bereau. *Unconstrained Structure Formation in Coarse-Grained Protein Simulations*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA, 2011.

[20] T. Bereau and M. Deserno. Generic coarse-grained model for protein folding and aggregation. *J. Chem. Phys.*, **130**:235106, 2009.

[21] T. Bereau, M. Deserno, and M. Bachmann. Structural basis of folding cooperativity in model proteins: Insights from a microcanonical perspective. *Biophys. J.*, **100**:2764, 2011.

[22] H. Berman, K. Henrick, and H. Nakamura. Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.*, **10**:980, 2003.

[23] H. Berman, K. Henrick, H. Nakamura, and J. L. Markley. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.*, **35**:D301, 2007.

[24] E. P. Bernard, W. Krauth, and D. B. Wilson. Event-chain Monte Carlo algorithms for hard-sphere systems. *Phys. Rev. E*, **80**:056704, 2009.

[25] A. Bhattacharyya, A. K. Thakur, V. M. Chellgren, G. Thiagarajan, A. D. Williams, B. W. Chellgren, T. P. Creamer, and R. Wetzel. Oligoproline effects on polyglutamine conformation and aggregation. *J. Mol. Biol.*, **355**:524, 2006.

[26] O. Bieri, J. Wirz, B. Hellrung, M. Schutkowski, M. Drewello, and T. Kiefhaber. The speed limit for protein folding measured by triplet-triplet energy transfer. *Proc. Natl. Acad. Sci. U. S. A.*, **96**:9597, 1999.

[27] F. Bogár, Z. Szekeres, F. Bartha, B. Penke, and J. Ladik. Density functional study of infinite polyserine chains. *Phys. Chem. Chem. Phys.*, **7**:2965, 2005.

[28] A. Böker. *Wang-Landau simulations of protein-like Gō model molecules*. Master's thesis, Martin-Luther-Universität Halle-Wittenberg, 2014.

[29] A. Böker and W. Paul. Wang-Landau simulation of Gō model molecules. *Eur. Phys. J. E*, **39**:5, 2016.

[30] M. Bonomi and M. Parrinello. Enhanced sampling in the well-tempered ensemble. *Phys. Rev. Lett.*, **104**:190601, 2010.

[31] W. L. Bragg, J. C. Kendrew, and M. F. Perutz. Polypeptide chain configurations in crystalline proteins. *Proc. Royal Soc. Lond.*, **A203**:321, 1950.

[32] C. Calero-Rubio, B. Paik, X. Jia, K. L. Kiick, and C. J. Roberts. Predicting unfolding thermodynamics and stable intermediates for alanine-rich helical peptides with the aid of coarse-grained molecular simulation. *Biophys. Chem.*, **217**:8, 2016.

[33] M. Carballo-Pacheco and B. Strodel. Comparison of force fields for Alzheimer's Aβ_{42}: A case study for intrinsically disordered proteins. *Protein Sci.*, **26**:174, 2017.

[34] A. Chakrabartty, A. J. Doig, and R. L. Baldwin. Helix capping propensities in peptides parallel those in proteins. *Proc. Natl. Acad. Sci. U. S. A.*, **90**:11332, 1993.

[35] Y. Chen and J. Ding. Construction of an intermediate-resolution lattice model and re-examination of the helix-coil transition: a dynamic Monte Carlo simulation. *J. Biomol. Struct. Dyn.*, **32**:792, 2014.

[36] M. Cheon, I. Chang, and C. K. Hall. Extending the PRIME model for protein aggregation to all 20 amino acids. *Proteins*, **78**:2950, 2010.

[37] M. Cheon, I. Chang, and C. K. Hall. Spontaneous formation of twisted Aβ_{16-22} fibrils in large-scale molecular-dynamics simulations. *Biophys. J.*, **101**:2493, 2011.

[38] M. Cheon, I. Chang, and C. K. Hall. Influence of temperature on formation of perfect tau fragment fibrils using PRIME20/DMD simulations. *Protein Sci.*, **21**:1514, 2012.

[39] M. Cheon, C. K. Hall, and I. Chang. Structural conversion of Aβ_{17-42} peptides from disordered oligomers to U-shape protofilaments via multiple kinetic pathways. *PLOS Comput. Biol.*, **11**:e1004258, 2015.

[40] M. Cheon, M. Kang, and I. Chang. Polymorphism of fibrillar structures depending on the size of assembled aβ_{17-42} peptides. *Sci. Rep.*, **6**:38196, 2016.

[41] S. L. Crick, M. Jayaraman, C. Frieden, R. Wetzel, and R. V. Pappu. Fluorescence correlation spectroscopy shows that monomeric polyglutamine molecules form collapsed structures in aqueous solutions. *Proc. Natl. Acad. Sci. U. S. A.*, **103**:16764, 2006.

[42] S. L. Crick, K. M. Ruff, K. Garai, C. Frieden, and R. V. Pappu. Unmasking the roles of N- and C-terminal flanking sequences from exon 1 of huntingtin as modulators of polyglutamine aggregation. *Proc. Natl. Acad. Sci. U. S. A.*, **110**:20075, 2013.

[43] M. Crisma, F. Formaggio, A. Moretto, and C. Toniolo. Peptide helices based on α-amino acids. *Biopolymers*, **84**:3, 2006.

[44] V. Daggett, P. A. Kollman, and I. D. Kuntz. A molecular dynamics simulation of polyalanine: An analysis of equilibrium motions and helix-coil transitions. *Biopolymers*, **31**:1115, 1991.

[45] J. E. Davies and D. C. Rubinsztein. Polyalanine and polyserine frameshift products in Huntington's disease. *J. Med. Genet.*, **43**:893, 2006.

[46] A. Davtyan, N. P. Schafer, W. Zheng, C. Clementi, P. G. Wolynes, and G. A. Papoian. AWSEM-MD: Protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing. *J. Phys. Chem. B*, **116**:8494, 2012.

[47] K. A. Dill. Theory for the folding and stability of globular proteins. *Biochemistry*, **24**:1501, 1985.

[48] K. A. Dill and H. S. Chan. From Levinthal to pathways to funnels. *Nat. Struct. Biol.*, **4**:10, 1997.

[49] K. A. Dill, S. B. Ozkan, M. S. Shell, and T. R. Weikl. The protein folding problem. *Annu. Rev. Biophys.*, **37**:289, 2008.

[50] F. Ding, J. M. Borreguero, S. V. Buldyrev, H. E. Stanley, and N. V. Dokholyan. Mechanism for the α-helix to β-hairpin transition. *Proteins*, **53**:220, 2003.

[51] C. M. Dobson, A. Šali, and M. Karplus. Protein folding: A perspective from theory and experiment. *Angew. Chem., Int. Ed.*, **37**:868, 1998.

[52] L. Dodd, T. Boone, and D. N. Theodorou. A concerted rotation algorithm for atomistic Monte Carlo simulation of polymer melts and glasses. *Mol. Phys.*, **78**:961, 1993.

[53] A. J. Doig. Recent advances in helix-coil theory. *Biophys. Chem.*, **101-102**:281, 2002.

[54] J. Donohue. Hydrogen bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. U. S. A.*, **39**:470, 1953.

[55] P. Doty, A. M. Holtzer, J. H. Bradbury, and E. R. Blout. Polypeptides. II. the configuration of polymers of γ-Benzyl-L-Glutamate in solution. *J. Am. Chem. Soc.*, **76**:4493, 1954.

[56] A. K. Dunker, C. J. Brown, J. D. Lawson, L. M. Iakoucheva, and Z. Obradović. Intrinsic disorder and protein function. *Biochemistry*, **41**:6573, 2002.

[57] M. Duyao, C. Ambrose, R. Myers, A. Novelletto, F. Persichetti, M. Frontali, S. Folstein, C. Ross, M. Franz, M. Abbott, J. Gray, P. Conneally, A. Young, J. Penney, Z. Hollingsworth, I. Shoulson, A. Lazzarini, A. Falek, W. Koroshetz, D. Sax, E. Bird, J. Vonsattel, E. Bonilla, J. Alvir, J. B. Conde, J.-H. Cha, L. Dure, F. Gomez, M. Ramos, J. Sanchez-Ramos, S. Snodgrass, M. de Young, N. Wexler, C. Moscowitz, G. Penchaszadeh, H. MacFarlande, M. Anderson, B. Jenkins, J. Srinidhi, G. Barnes, J. Gusella, and M. MacDonald. Trinucleotide repeat length instability and age of onset in Huntington's disease. *Nat. Genet.*, **4**:387, 1993.

[58] H. J. Dyson and P. E. Wright. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.*, **6**:197, 2005.

[59] S. F. Edwards. The statistical mechanics of polymers with excluded volume. *Proc. Phys. Soc.*, **85**:613, 1965.

[60] D. Eisenberg and M. Jucker. The amyloid state of proteins in human diseases. *Cell*, **148**:1188, 2012.

[61] P. Enke. Effect of different crowding agents on structure and dynamics of unfolded proteins. In *27$^{\text{th}}$ Faltertage on "Protein Folding, Dynamics and Stability"*, Oct 2016.

[62] P. Enke, M. Schleeger, and T. Kiefhaber. Dynamic and structural properties of polyglutamine. In *International Discussion Meeting on Polymer Crystallization 2017*, Sep 2017.

[63] B. Fierz, A. Reiner, and T. Kiefhaber. Local conformational dynamics in α-helices measured by fast triplet transfer. *Proc. Natl. Acad. Sci. U. S. A.*, **106**:1057, 2009.

[64] B. Fierz, H. Satzger, C. Root, P. Gilch, W. Zinth, and T. Kiefhaber. Loop formation in unfolded polypeptide chains on the picoseconds to microseconds time scale. *Proc. Natl. Acad. Sci. U. S. A.*, **104**:2163, 2007.

[65] N. C. Fitzkee and G. D. Rose. Steric restrictions in protein folding: An α-helix cannot be followed by a contiguous β-strand. *Protein Sci.*, **13**:633, 2004.

[66] P. J. Fleming and G. D. Rose. Do all backbone polar groups in proteins form hydrogen bonds? *Protein Sci.*, **14**:1911, 2005.

[67] A. M. Fluitt and J. J. de Pablo. An analysis of biomolecular force fields for simulations of polyglutamine in solution. *Biophys. J.*, **109**:1009, 2015.

[68] D. Frenkel. Introduction to Monte Carlo methods. In N. Attig, K. Binder, H. Grubmüller, and K. Kremer, editors, *Computational Soft Matter: From Synthetic Polymers to Proteins*, chapter 2, page 29. John von Neumann Institute for Computing, 2004.

[69] D. Frishman and P. Argos. Knowledge-based protein secondary structure assignment. *Proteins*, **23**:566, 1995.

[70] B. C. Gin, J. P. Garrahan, and P. L. Geissler. The limited role of nonnative contacts in the folding pathways of a lattice protein. *J. Mol. Biol.*, **392**:1303, 2009.

[71] S. Gnanakaran and A. E. García. Helix-coil transition of alanine peptides in water: Force field dependence on the folded and unfolded structures. *Proteins*, **59**:773, 2005.

[72] S. M. Gopal, S. Mukherjee, Y.-M. Cheng, and M. Feig. PRIMO/PRIMONA: A coarse-grained model for proteins and nucleic acids that preserves near-atomistic accuracy. *Proteins*, **78**:1266, 2010.

[73] M. Gopalswamy, A. Kumar, J. Adler, M. Baumann, M. Henze, S. T. Kumar, M. Fändrich, H. A. Scheidt, D. Huster, and J. Balbach. Structural characterization of amyloid fibrils from the human parathyroid hormone. *Biochim. Biophys. Acta - Proteins and Proteomics*, **1854**:249, 2015.

[74] C. K. Hall and V. A. Wagoner. Computational approaches to fibril structure and formation. In *Methods Enzymol.*, page 338. Elsevier, 2006.

[75] U. H. E. Hansmann and Y. Okamoto. Finite-size scaling of helix-coil transitions in poly-alanine studied by multicanonical simulations. *J. Chem. Phys.*, **110**:1267, 1999.

[76] J. A. Hardy and G. A. Higgins. Alzheimer's disease: the amyloid cascade hypothesis. *Science*, **256**:184, 1992.

[77] H. Havukainen, J. Underhaug, F. Wolschin, G. Amdam, and Ø. Halskau. A vitellogenin polyserine cleavage site: highly disordered conformation protected from proteolysis by phosphorylation. *J. Exp. Biol.*, **215**:1837, 2012.

[78] S. Hilbert, P. Hänggi, and J. Dunkel. Thermodynamic laws in isolated systems. *Phys. Rev. E*, **90**:062116, 2014.

[79] R. D. Hills Jr. and C. L. Brooks III. Insights from coarse-grained G$\overline{\text{o}}$ models for protein folding and dynamics. *Int. J. Mol. Sci.*, **10**:889, 2009.

[80] B. K. Ho, A. Thomas, and R. Brasseur. Revisiting the Ramachandran plot: Hard-sphere repulsion, electrostatics, and H-bonding in the α-helix. *Protein Sci.*, **12**:2508, 2003.

[81] B. Honig and A.-S. Yang. Free energy balance in protein folding. *Adv. Protein Chem.*, **46**:27, 1995.

[82] M. B. Howard, N. A. Ekborg, L. E. Taylor, S. W. Hutcheson, and R. M. Weiner. Identification and analysis of polyserine linker domains in prokaryotic proteins with emphasis on the marine bacterium Microbulbifer degradans. *Protein Sci.*, **13**:1422, 2004.

[83] R. R. Hudgins and M. F. Jarrold. Helix formation in unsolvated alanine-based peptides: Helical monomers and helical dimers. *J. Am. Chem. Soc.*, **121**:3494, 1999.

[84] G. Huntington. On chorea. *The Medical and Surgical Reporter*, **26**:317, 1872.

[85] M. A. Huntley and G. B. Golding. Evolution of simple sequence in proteins. *J. Mol. Evol.*, **51**:131, 2000.

[86] M. A. Huntley and G. B. Golding. Selection and slippage creating serine homopolymers. *Mol. Biol. Evol.*, **23**:2017, 2006.

[87] A. Irbäck, F. Sjunnesson, and S. Wallin. Three-helix-bundle protein in a Ramachandran model. *Proc. Natl. Acad. Sci. U. S. A.*, **97**:13614, 2000.

[88] W. Janke and W. Paul. Thermodynamics and structure of macromolecules from flat-histogram Monte Carlo simulations. *Soft Matter*, **12**:642, 2016.

[89] L. A. Johnson, A. Monge, and R. A. Friesner. A hierarchical algorithm for polymer simulations. *J. Chem. Phys.*, **97**:9355, 1992.

[90] W. Kabsch and C. Sander. Dictionary of Protein Secondary Structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**:2577, 1983.

[91] S. C. Kapfer and W. Krauth. Irreversible local markov chains with rapid convergence towards equilibrium. *Phys. Rev. Lett.*, **119**:240603, 2017.

[92] A. Kapoor and A. Travesset. Folding and stability of helical bundle proteins from coarse-grained models. *Proteins*, **81**:1200, 2013.

[93] K. Kar, C. L. Hoop, K. W. Drombosky, M. A. Baker, R. Kodali, I. Arduini, P. C. A. van der Wel, W. S. Horne, and R. Wetzel. β-hairpin-mediated nucleation of polyglutamine amyloid formation. *J. Mol. Biol.*, **425**:1183, 2013.

[94] K. Kar, M. Jayaraman, B. Sahoo, R. Kodali, and R. Wetzel. Critical nucleus size for disease-related polyglutamine aggregation is repeat-length dependent. *Nat. Struct. Mol. Biol.*, **18**:328, 2011.

[95] N. C. Karayiannis, A. E. Giannousaki, V. G. Mavrantzas, and D. N. Theodorou. Atomistic Monte Carlo simulation of strictly monodisperse long polyethylene melts through a generalized chain bridging algorithm. *J. Chem. Phys.*, **117**:5465, 2002.

[96] P. Käthner. *Simulation von Poly-Serin mittels Stochastic Approximation Monte Carlo*. Bachelor's thesis, Martin-Luther-Universität Halle-Wittenberg, 2017.

[97] H. G. Katzgraber. Introduction to monte carlo methods, 2011. Lecture at the third international summer school "Modern Computation Science", 15 - 26 August 2011, Oldenburg (Germany).

[98] K. Kieburtz, M. MacDonald, C. Shih, A. Feigin, K. Steinberg, K. Bordwell, C. Zimmerman, J. Srinidhi, J. Sotack, J. Gusella, and I. Shoulson. Trinucleotide repeat length and progression of illness in Huntington's disease. *J. Med. Genet.*, **31**:872, 1994.

[99] M. W. Kim, Y. Chelliah, S. W. Kim, Z. Otwinowski, and I. Bezprozvanny. Secondary structure of huntingtin amino-terminal region. *Structure*, **17**:1205, 2009.

[100] B. S. Kinnear, D. T. Kaleta, M. Kohtani, R. R. Hudgins, and M. F. Jarrold. Conformations of unsolvated valine-based peptides. *J. Am. Chem. Soc.*, **122**:9243, 2000.

[101] T. P. J. Knowles, M. Vendruscolo, and C. M. Dobson. The amyloid state and its association with protein misfolding diseases. *Nat. Rev. Mol. Cell Biol.*, **15**:384, 2014.

[102] A. Kornberg, L. L. Bertsch, J. F. Jackson, and H. G. Khorana. Enzymatic synthesis of deoxyribonucleic acid, XVI. oligonucleotides as templates and the mechanism of their replication. *Proc. Natl. Acad. Sci. U. S. A.*, **51**:315, 1964.

[103] F. Krieger, B. Fierz, F. Axthelm, K. Joder, D. Meyer, and T. Kiefhaber. Intrachain diffusion in a protein loop fragment from carp parvalbumin. *Chem. Phys.*, **307**:209, 2004.

[104] F. Krieger, B. Fierz, O. Bieri, M. Drewello, and T. Kiefhaber. Dynamics of unfolded polypeptide chains as model for the earliest steps in protein folding. *J. Mol. Biol.*, **332**:265, 2003.

[105] R. Laghaei and N. Mousseau. Spontaneous formation of polyglutamine nanotubes with molecular dynamics simulations. *J. Chem. Phys.*, **132**:165102, 2010.

[106] V. V. Lakhani, F. Ding, and N. V. Dokholyan. Polyglutamine induced misfolding of huntingtin exon1 is modulated by the flanking sequences. *PLOS Comput. Biol.*, **6**:e1000772, 2010.

[107] D. C. Latshaw. *Effects of Crowders, Inhibitors, and Interfaces on Peptide Aggregation Using Coarse-Grained Simulations*. PhD thesis, North Carolina State University, 2015.

[108] D. C. Latshaw, M. Cheon, and C. K. Hall. Effects of macromolecular crowding on amyloid beta (16–22) aggregation using coarse-grained simulations. *J. Phys. Chem. B*, **118**:13513, 2014.

[109] D. C. Latshaw and C. K. Hall. Effects of hydrophobic macromolecular crowders on amyloid β (16-22) aggregation. *Biophys. J.*, **109**:124, 2015.

[110] G. Levinson and G. A. Gutman. Slipped-strand mispairing: A major mechanism for DNA and sequence evolution. *Mol. Biol. Evol.*, **4**:203, 1987.

[111] C. Levinthal. Are there pathways for protein folding? *J. Chim. Phys.*, **65**:44, 1968.

[112] C. Levinthal. How to fold graciously. In *Mössbauer Spectroscopy in Biological Systems*, 1969.

[113] F. Liang. A theory on Flat Histogram Monte Carlo algorithms. *J. Stat. Phys.*, **122**:511, 2006.

[114] F. Liang. An overview of Stochastic Approximation Monte Carlo. *WIREs Comput. Stat.*, **6**:240, 2014.

[115] F. Liang, C. Liu, and R. J. Carroll. Stochastic approximation in Monte Carlo computation. *J. Am. Stat. Assoc.*, **102**:305, 2007.

[116] Y. Lin and J. K. Gross. Molecular cloning and characterization of winter flounder antifreeze cDNA. *Proc. Natl. Acad. Sci. U. S. A.*, **78**:2825, 1981.

[117] H. A. Lorentz. Ueber die Anwendung des Satzes vom Virial in der kinetischen Theorie der Gase. *Ann. Phys. (Berl.)*, **248**:127, 1881.

[118] P. C. Lyu, M. I. Liff, L. A. Marky, and N. R. Kallenbach. Side chain contributions to the stability of alpha-helical structure in peptides. *Science*, **250**:669, 1990.

[119] J. E. Magee, Z. Song, R. A. Curtis, and L. Lue. Structure and aggregation of a helix-forming polymer. *J. Chem. Phys.*, **126**:144911, 2007.

[120] G. I. Makhatadze and P. L. Privalov. Contribution of hydration to protein folding thermodynamics I. The enthalpy of hydration. *J. Mol. Biol.*, **232**:639, 1993.

[121] A. J. Marchut. *Simulation of polyglutamine aggregation with an intermediate resolution protein model*. PhD thesis, North Carolina State University, 2006.

[122] A. J. Marchut and C. K. Hall. Side-chain interactions determine amyloid formation by model polyglutamine peptides in molecular dynamics simulations. *Biophys. J.*, **90**:4574, 2006.

[123] A. J. Marchut and C. K. Hall. Spontaneous formation of annular structures observed in molecular dynamics simulations of polyglutamine peptides. *Comput. Biol. Chem.*, **30**:215, 2006.

[124] A. J. Marchut and C. K. Hall. Effects of chain length on the aggregation of model polyglutamine peptides: Molecular dynamics simulations. *Proteins*, **66**:96, 2007.

[125] S. Marqusee, V. H. Robbins, and R. L. Baldwin. Unusually stable helix formation in short alanine-based peptides. *Proc. Natl. Acad. Sci. U. S. A.*, **86**:2586, 1989.

[126] K. Marx and F. Engels. *Manifesto of the Communist Party*. 1848. English translation by Samuel Moore (1888). Url: https://www.marxists.org/archive/marx/works/1848/communist-manifesto/index.htm, last visited 2019/02/18.

[127] V. G. Mavrantzas, T. D. Boone, E. Zervopoulou, and D. N. Theodorou. End-bridging Monte Carlo: A fast algorithm for atomistic simulation of condensed phases of long polymer chains. *Macromolecules*, **32**:5072, 1999.

[128] D. Mazzier, L. Grassi, A. Moretto, C. Alemán, F. Formaggio, C. Toniolo, and M. Crisma. En route towards the peptide γ-helix: X-ray diffraction analyses and conformational energy calculations of Adm-rich short peptides. *J. Pept. Sci.*, **23**:346, 2016.

[129] R. P. Menon and A. Pastore. Expansion of amino acid homo-sequences in proteins: Insights into the role of amino acid homo-polymers and of the protein context in aggregation. *Cell. Mol. Life Sci.*, **63**:1677, 2006.

[130] C. Messaed and G. A. Rouleau. Molecular mechanisms underlying polyalanine diseases. *Neurobiol. Dis.*, **34**:397, 2009.

[131] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, **21**:1087, 1953.

[132] J. B. O. Mitchell and S. L. Price. The nature of the n–h$\cdots$o=c hydrogen bond: An intermolecular perturbation theory study of the formamide/formaldehyde complex. *J. Comp. Chem*, **11**:1217, 1990.

[133] S. Miyazawa and R. L. Jernigan. Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. *Macromolecules*, **18**:534, 1985.

[134] S. Miyazawa and R. L. Jernigan. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.*, **256**:623, 1996.

[135] A. Möglich, F. Krieger, and T. Kiefhaber. Molecular basis for the effect of urea and guanidinium chloride on the dynamics of unfolded polypeptide chains. *J. Mol. Biol.*, **345**:153, 2005.

[136] G. E. Moore. Cramming more components onto integrated circuits. *Electronics*, **38**:114, 1965.

[137] Y. Mu and M. Yu. Effects of hydrophobic interaction strength on the self-assembled structures of model peptides. *Soft Matter*, **10**:4956, 2014.

[138] S. Neumaier, A. Reiner, M. Büttner, B. Fierz, and T. Kiefhaber. Testing the diffusing boundary model for the helix-coil transition in peptides. *Proc. Natl. Acad. Sci. U. S. A.*, **110**:12905, 2013.

[139] H. D. Nguyen. *Computer simulations of protein folding and aggregation*. PhD thesis, North Carolina State University, 2004.

[140] H. D. Nguyen and C. K. Hall. Molecular dynamics simulations of spontaneous fibril formation by random-coil peptides. *Proc. Natl. Acad. Sci. U. S. A.*, **101**:16180, 2004.

[141] H. D. Nguyen and C. K. Hall. Phase diagrams describing fibrillization by polyalanine peptides. *Biophys. J.*, **87**:4122, 2004.

[142] H. D. Nguyen and C. K. Hall. Kinetics of fibril formation by polyalanine peptides. *J. Biol. Chem.*, **280**:9074, 2005.

[143] H. D. Nguyen and C. K. Hall. Spontaneous fibril formation by polyalanines; discontinuous molecular dynamics simulations. *J. Am. Chem. Soc.*, **128**:1890, 2006.

[144] H. D. Nguyen, A. J. Marchut, and C. K. Hall. Solvent effects on the conformational transition of a model polyalanine peptide. *Protein Sci.*, **13**:2909, 2004.

[145] M. Oka, Y. Baba, A. Kagemoto, and A. Nakajima. γ-helix; new-type helical conformation in proteins found through theoretical analysis on elastin-model polypeptide. *Polym. J.*, **22**:555, 1990.

[146] Y. Okamoto and U. H. E. Hansmann. Thermodynamics of helix-coil transitions studied by multicanonical algorithms. *J. Chem. Phys.*, **99**:11276, 1995.

[147] K. O'Neil and W. F. DeGrado. A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids. *Science*, **250**:646, 1990.

[148] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes. Theory of protein folding: The energy landscape perspective. *Annu. Rev. Phys. Chem.*, **48**:545, 1997.

[149] P. Palenčár and T. Bleha. Folding of polyalanine into helical hairpins. *Macromol. Theory Simul.*, **19**:488, 2010.

[150] P. Palenčár and T. Bleha. Molecular dynamics simulations of the folding of poly(alanine) peptides. *J. Mol. Model.*, **17**:2367, 2011.

[151] K. M. Pan, M. Baldwin, J. Nguyen, M. Gasset, A. Serban, D. Groth, I. Mehlhorn, Z. Huang, R. J. Fletterick, and F. E. Cohen. Conversion of alpha-helices into beta-sheets features in the formation of the scrapie prion proteins. *Proc. Natl. Acad. Sci. U. S. A.*, **90**:10962, 1993.

[152] N. Panasik, P. J. Fleming, and G. D. Rose. Hydrogen-bonded turns in proteins: The case for a recount. *Protein Sci.*, **14**:2910, 2005.

[153] G. B. Panigrahi, R. Lau, S. E. Montgomery, M. R. Leonard, J. L. Marcadier, M. Kekis, C. Vosch, A. Todd, and C. E. Pearson. Error-prone repair of slipped (CTG)·(CAG) repeats and disease-associated expansions. In R. D. Wells and T. Ashizawa, editors, *Genetic Instabilities and Neurological Diseases*, chapter 33, page 487. Academic Press, second edition, 2006.

[154] P. V. K. Pant and D. N. Theodorou. Variable connectivity method for the atomistic Monte Carlo simulation of polydisperse polymer melts. *Macromolecules*, **28**:7224, 1995.

[155] T. Paracelsus. *von den kranckheyten so die vernunfft berauben*. Basel, 1567. pp. 31ff.

[156] L. Pauling and R. B. Corey. Atomic coordinates and structure factors for two helical configurations of polypeptide chains. *Proc. Natl. Acad. Sci. U. S. A.*, **37**:235, 1951.

[157] L. Pauling and R. B. Corey. The pleated sheet, a new layer configuration of polypeptide chains. *Proc. Natl. Acad. Sci. U. S. A.*, **37**:251, 1951.

[158] L. Pauling, R. B. Corey, and H. R. Branson. The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. U. S. A.*, **37**:205, 1951.

[159] C. E. Pearson and R. R. Sinden. Trinucleotide repeat DNA structures: dynamic mutations from dynamic DNA. *Curr. Opin. Struct. Biol.*, **8**:321, 1998.

[160] M. F. Perutz. New X-ray evidence on the configuration of polypeptide chains: Polypeptide chains in poly-γ-benzyl-L-glutamate, keratin and hæmoglobin. *Nature*, **167**:1053, 1951.

[161] M. F. Perutz, T. Johnson, M. Suzuki, and J. T. Finch. Glutamine repeats as polar zippers: Their possible role in inherited neurodegenerative diseases. *Proc. Natl. Acad. Sci. U. S. A.*, **91**:5355, 1994.

[162] M. F. Perutz and A. H. Windle. Cause of neural death in neurodegenerative diseases attributable to expansion of glutamine repeats. *Nature*, **412**:143, 2001.

[163] E. M. Phelps. *Polyalanine and Aβ Aggregation Kinetics: Probing Intermediate Oligomer Formation and Structure Using Computer Simulations*. PhD thesis, North Carolina State University, 2011.

[164] E. M. Phelps and C. K. Hall. Structural transitions and oligomerization along polyalanine fibril formation pathways from computer simulations. *Proteins*, **80**:1582, 2012.

[165] H. Prasad and S. Singh. Existence of γ-helix in natural proteins. *Int. J. Biol. Macromol.*, **3**:243, 1980.

[166] S. Pylaeva, A. Böker, H. Elgabarty, W. Paul, and D. Sebastiani. The conformational ensemble of polyglutamine-14 chains: specific influence of solubility tail and chromophores. *ChemPhysChem*, **19**:2931, 2018.

[167] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. Url: https://www.R-project.org/.

[168] G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.*, **7**:95, 1963.

[169] F. Rampf, K. Binder, and W. Paul. The phase diagram of a single polymer chain: New insights from a new simulation method. *J. Pol. Sci. B*, **44**:2542, 2006.

[170] N. Rathore and J. J. de Pablo. Monte Carlo simulation of proteins through a random walk in energy space. *J. Chem. Phys.*, **116**:7225, 2002.

[171] P. Ricchiuto, A. V. Brukhno, and S. Auer. Protein aggregation: Kinetics versus thermodynamics. *J. Phys. Chem. B*, **116**:5384, 2012.

[172] C. A. Rohl, W. Fiori, and R. L. Baldwin. Alanine is helix-stabilizing in both template-nucleated and standard peptide helices. *Proc. Natl. Acad. Sci. U. S. A.*, **96**:3682, 1999.

[173] M. N. Rosenbluth and A. W. Rosenbluth. Monte Carlo calculation of the average extension of molecular chains. *J. Chem. Phys.*, **23**:356, 1955.

[174] G. O. Rutter. *Coarse-Grained Simulations of Intrinsically Disordered Peptides*. Phd thesis, University of Warwick, Oct 2015.

[175] G. O. Rutter, A. H. Brown, D. Quigley, T. R. Walsh, and M. P. Allen. Testing the transfer of a coarse-grained model to intrinsically disordered proteins. *Phys. Chem. Chem. Phys.*, **17**:31741, 2015.

[176] J. M. Scholtz, H. Qian, E. J. York, J. M. Stewart, and R. L. Baldwin. Parameters of helix-coil transition theory for alanine-based peptides of varying chain lengths in water. *Biopolymers*, **31**:1463, 1991.

[177] Schrödinger, LLC. *The PyMOL Molecular Graphics System, Version 1.7.0.0*, 2014. Url: https://pymol.org/.

[178] T. Shakirov. Convergence estimation of flat-histogram algorithms based on simulation results. *Comput. Phys. Commun.*, **228**:38, 2018.

[179] T. Shakirov, S. Zablotskiy, A. Böker, V. A. Ivanov, and W. Paul. Comparison of Boltzmann and Gibbs entropies for the analysis of single-chain phase transitions. *Eur. Phys. J. ST*, **226**:705, 2017.

[180] Q. Shao and C. K. Hall. A discontinuous potential model for protein-protein interactions. In R. Q. Snurr, C. S. Adjiman, and D. A. Kofke, editors, *Foundations of Molecular Modeling and Simulation*, page 1. Springer, 2016.

[181] S.-Y. Sheu, D.-Y. Yang, H. L. Selzle, and E. W. Schlag. Energetics of hydrogen bonds in peptides. *Proc. Natl. Acad. Sci. U. S. A.*, **100**:12683, 2003.

[182] R. G. Snell, J. C. MacMillan, J. P. Cheadle, I. Fenton, L. P. Lazarou, P. Davies, M. E. MacDonald, J. F. Gusella, P. S. Harper, and D. J. Shaw. Relationship between trinucleotide repeat expansion and phenotypic variation in Huntington's disease. *Nat. Genet.*, **4**:393, 1993.

[183] S. Sudo, T. Fujikawa, T. Nagakura, T. Ohkubo, K. Sakaguchi, M. Tanaka, K. Nakashima, and T. Takahashi. Structures of mollusc shell framework proteins. *Nature*, **387**:563, 1997.

[184] S. Sun. Reduced representation model of protein structure prediction: Statistical potential and genetic algorithms. *Protein Sci.*, **2**:762, 1993.

[185] D. Suvlu, S. Samaratunga, D. Thirumalai, and J. C. Rasaiah. Thermodynamics of helix-coil transitions of polyalanine in open carbon nanotubes. *J. Phys. Chem. Lett.*, **8**:494, 2017.

[186] S. Tabrizi, B. Leavitt, H. Kordasiewicz, C. Czech, E. Swayze, D. A. Norris, T. Baumann, I. Gerlach, S. Schobel, A. Smith, R. Lane, and C. F. Bennett. Effects of IONIS-HTTRx in patients with early Huntington's disease, results of the first HTT-lowering drug trial (CT.002). *Neurology*, **90**(15 Supplement), 2018.

[187] S. Takada, Z. Luthey-Schulten, and P. G. Wolynes. Folding dynamics with nonadditive forces: A simulation study of a designed helical protein and a random heteropolymer. *J. Chem. Phys.*, **110**:11616, 1999.

[188] M. Takano, T. Yamato, J. Higo, A. Suyama, and K. Nagayama. Molecular dynamics of a 15-residue poly(L-alanine) in water: Helix formation and energetics. *J. Am. Chem. Soc.*, **121**:605, 1999.

[189] H. Taketomi, Y. Ueda, and N. Gō. Studies on protein folding, unfolding and fluctuations by computer simulation I. the effect of specific amino acid sequence represented by specific inter-unit interactions. *Int. J. Pept. Protein Res.*, **7**:445, 1975.

[190] P. Tamiozzo, P. M. A. Lucchesi, and A. Ambrogi. Monitoring for Mycoplasma hyopneumoniae before and after a partial depopulation program using a typing scheme based on the polyserine repeat motif of p146. *Journal of Swine Health and Production*, **21**:309, 2013.

[191] M. P. Taylor, W. Paul, and K. Binder. All-or-none protein folding transition of a flexible homopolymer chain. *Phys. Rev. E*, **79**:050801, 2009.

[192] M. P. Taylor, W. Paul, and K. Binder. Applications of the Wang-Landau algorithm to phase transitions of a single polymer chain. *Polym. Sci. Ser. C*, **55**:23, 2013.

[193] The Huntington's Disease Collaborative Research Group. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's Disease chromosomes. *Cell*, **72**:971, 1993.

[194] D. N. Theodorou and U. W. Suter. Shape of unperturbed linear polymers: polypropylene. *Macromolecules*, **18**:1206, 1985.

[195] P. A. Thompson, W. A. Eaton, and J. Hofrichter. Laser temperature jump study of the helix⇌coil kinetics of an alanine peptide interpreted with a 'kinetic zipper' model. *Biochemistry*, **36**:9200, 1997.

[196] G. M. Torrie and J. P. Valleau. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comp. Phys.*, **23**:187, 1977.

[197] V. Tozzini. Minimalist models for proteins: a comparative analysis. *Q. Rev. Biophys.*, **43**:333, 2010.

[198] K. Tsuji, H. Ohe, and H. Watanabe. Stability of the PBLG helix at high temperature and the wormlike character of the helix. *Polym. J.*, **4**:553, 1973.

[199] B. Urbanc, J. M. Borreguero, L. Cruz, and H. E. Stanley. Ab initio discrete molecular dynamics approach to protein folding and aggregation. *Methods Enzymol.*, **412**:314, 2006.

[200] C. M. Venkatachalam. Stereochemical criteria for polypeptides and proteins. V. conformation of a system of three linked peptide units. *Biopolymers*, **6**:1425, 1968.

[201] A. Vitalis, N. Lyle, and R. V. Pappu. Thermodynamics of β-sheet formation in polyglutamine. *Biophys. J.*, **97**:303, 2009.

[202] A. Vitalis, X. Wang, and R. V. Pappu. Atomistic simulations of the effect of polyglutamine chain length and solvent quality on conformational equilibria and spontaneous homodimerization. *J. Mol. Biol.*, **384**:279, 2008.

[203] A. Voegler Smith. *Simulations of Protein Refolding and Aggregation Using a Novel Intermediate-Resolution Protein Model*. PhD thesis, North Carolina State University, 2001.

[204] A. Voegler Smith and C. K. Hall. Bridging the gap between homopolymer and protein models: A discontinuous molecular dynamics study. *J. Chem. Phys.*, **113**:9331, 2000.

[205] A. Voegler Smith and C. K. Hall. α-helix formation: Discontinuous molecular dynamics on an intermediate-resolution protein model. *Proteins*, **44**:344, 2001.

[206] A. Voegler Smith and C. K. Hall. Assembly of a tetrameric α-helical bundle: Computer simulations on an intermediate-resolution protein model. *Proteins*, **44**:376, 2001.

[207] A. Voegler Smith and C. K. Hall. Protein refolding versus aggregation: computer simulations on an intermediate-resolution protein model. *J. Mol. Biol.*, **312**:187, 2001.

[208] J. Vymětal and J. Vondrášek. Gyration-and inertia-tensor-based collective coordinates for metadynamics. Application on the conformational behavior of polyalanine peptides and Trp-cage folding. *J. Phys. Chem. A*, **115**:11455, 2011.

[209] J. Wägele, S. D. Sio, B. Voigt, J. Balbach, and M. Ott. How fluorescent tags modify oligomer size distributions of the Alzheimer peptide. *Biophys. J.*, **116**:227, 2019.

[210] V. A. Wagoner. *Computer Simulation Studies of Self-Assembly of Fibril-Forming Peptides with an Intermediate Resolution Protein Model*. PhD thesis, North Carolina State University, 2010.

[211] V. A. Wagoner, M. Cheon, I. Chang, and C. K. Hall. Computer simulation study of amyloid fibril formation by palindromic sequences in prion peptides. *Proteins*, **79**:2132, 2011.

[212] V. A. Wagoner, M. Cheon, I. Chang, and C. K. Hall. Fibrillization propensity for short designed hexapeptides predicted by computer simulation. *J. Mol. Biol.*, **416**:598, 2012.

[213] V. A. Wagoner, M. Cheon, I. Chang, and C. K. Hall. Impact of sequence on the molecular assembly of short amyloid peptides. *Proteins*, **82**:1469, 2014.

[214] R. H. Walters and R. M. Murphy. Examining polyglutamine peptide length: A connection between collapsed conformations and increased aggregation. *J. Mol. Biol.*, **393**:978, 2009.

[215] F. Wang and D. P. Landau. Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys. Rev. Lett.*, **86**:2050, 2001.

[216] Y. Wang. *Understanding In Vitro Amyloid Formation and Inhibition Using Coarse-Grained Simulations.* PhD thesis, North Carolina State University, 2018.

[217] Y. Wang, Y. Gao, S. E. Hill, D. J. E. Huard, M. O. Tomlin, R. L. Lieberman, A. K. Paravastu, and C. K. Hall. Simulations and experiments delineate amyloid fibrilization by peptides derived from glaucoma-associated myocilin. *J. Phys. Chem. B*, **122**:5845, 2018.

[218] Y. Wang and C. K. Hall. Seeding and cross-seeding fibrillation of N-terminal prion protein peptides PrP(120-144). *Protein Sci.*, **27**:1304, 2018.

[219] Y. Wang, Q. Shao, and C. K. Hall. N-terminal prion protein peptides (PrP(120-144)) form parallel in-register β-sheets via multiple nucleation-dependent pathways. *J. Biol. Chem.*, **291**:22093, 2016.

[220] J. D. Watson and F. Crick. Molecular structure of nucleic acids – a structure for deoxyribose nucleic acid. *Nature*, **171**:737, 1953.

[221] Y. Wei, W. Nadler, and U. H. E. Hansmann. On the helix-coil transition in alanine based polypeptides in gas phase. *J. Chem. Phys.*, **126**:204307, 2007.

[222] B. Werlich. *Intramolekulare Strukturbildung durch Steifigkeitsvariation und Auswirkung der Anwesenheit von spezifischen Wechselwirkungen*. PhD thesis, MLU Halle-Wittenberg, 2017.

[223] B. Werlich, T. Shakirov, M. P. Taylor, and W. Paul. Stochastic approximation Monte Carlo and Wang-Landau Monte Carlo applied to a continuum polymer model. *Comput. Phys. Commun.*, **186**:65, 2015.

[224] B. Werlich, M. P. Taylor, and W. Paul. Wang-Landau and Stochastic Approximation Monte Carlo for semi-flexible polymer chains. *Physics Procedia*, **57**:82, 2014.

[225] R. Wetzel. Physical chemistry of polyglutamine: Intriguing tales of a monotonous sequence. *J. Mol. Biol.*, **421**:466, 2012.

[226] H. Wu, P. G. Wolynes, and G. A. Papoian. AWSEM-IDP: A coarse-grained force field for intrinsically disordered proteins. *J. Phys. Chem. B*, **122**:11115, 2018.

[227] wwPDB consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.*, **47**:D520, 2018.

[228] A.-S. Yang and B. Honig. Free energy determinants of secondary structure formation: I. α-helices. *J. Mol. Biol.*, **252**:351, 1995.

[229] Y. Zhang, S. Qiu, S. Jia, D. Xu, T. Ran, and W. Wang. Crystal structure of the sensor domain of BaeS from Serratia marcescens FS14. *Proteins*, **85**:1784, 2017.

[230] U. G. Zinth. *End-to-End Distance Distribution and Intra-Chain Diffusion in Unfolded Polypeptide Chains Determined by Time-Resolved FRET Measurements*. PhD thesis, TU München, 2014.

[231] C. Zuccato, M. Valenza, and E. Cattaneo. Molecular mechanisms and potential therapeutical targets in Huntington's disease. *Physiol. Rev*, **90**:905, 2010.

Appendix A

PRIME20 parameters of all amino acids

The tables of model parameters in chapter 2 only include the five amino acids which occurred in the simulations performed thus far. Here they are listed for all twenty proteinogenic amino acids.

Table A.1 – Side chain energies (table 2.3). Positive values are coloured blue, negative values red. The side chains are grouped together as they are in PRIME20.

	LIV	M	F	Y	W	H	ST	NQ	DE	KR	C	P	A
LIV	.200	.200	.203	.203	.203	.015	.015	.015	.015	.015	.139	.015	.148
M		.200	.203	.210	.210	.116	.116	.116	.015	.116	.139	.015	.148
F			.205	.205	.205	.015	.015	.015	.015	.015	.139	.015	.148
Y				.201	.201	.086	.086	.086	.086	.086	.116	.015	.148
W					.205	.086	.116	.086	.086	.015	.116	.015	.148
H						.080	.086	.080	.086	.086	.116	.074	.074
ST							.086	.086	.086	.086	.116	.074	.074
NQ								.080	.086	.086	.116	.074	.074
DE									.253	.136	.116	.074	.074
KR										.073	.116	.074	.074
C											.585	.015	.139
P												.074	.074
A													.084

Table A.2 – Side chain bead diameters in Å (table 2.4, left side)

	L	I	V	M	F	Y	W	H	S	T	N	Q	D	E	K	R	C	P	A
L	3.4	3.4	3.0	3.6	3.4	3.2	3.4	3.2	3.0	3.2	3.4	3.5	3.0	3.3	3.5	3.4	3.4	3.5	2.7
I		3.3	3.3	3.6	3.4	3.0	3.2	3.1	2.6	3.0	2.8	3.1	3.4	3.2	2.9	3.6	3.3	3.5	2.9
V			3.3	3.0	3.2	3.0	2.9	3.1	2.8	2.8	3.1	3.3	3.0	3.1	3.1	3.1	2.9	3.3	2.7
M				3.7	3.2	3.2	3.2	3.6	3.2	3.6	3.5	3.4	3.6	3.3	3.7	3.7	3.4	3.7	2.9
F					3.3	3.2	3.4	2.9	2.9	2.8	2.7	3.3	3.1	3.3	3.5	3.3	3.2	3.1	2.4
Y						3.0	3.2	3.1	2.9	3.2	3.3	3.4	2.8	3.3	3.5	3.1	2.9	3.3	2.7
W							3.7	3.2	2.7	3.3	2.8	3.4	3.2	3.5	3.5	3.0	3.3	3.4	2.7
H								3.4	2.6	2.9	3.4	3.3	2.8	3.3	3.4	3.5	2.8	3.7	3.1
S									2.5	2.9	3.0	2.7	2.8	2.9	3.0	3.0	2.8	3.2	2.3
T										2.9	3.1	3.3	3.1	3.1	3.1	3.2	2.7	2.6	2.6
N											3.3	3.5	3.2	3.1	3.2	2.9	3.1	3.3	2.8
Q												3.6	2.8	2.9	3.4	3.6	3.1	3.6	3.0
D													3.4	2.9	3.0	3.0	3.2	3.2	2.6
E														3.2	3.4	3.1	2.7	3.5	2.9
K															3.5	3.9	2.7	3.6	3.3
R																3.2	3.3	3.0	3.0
C																	2.1	3.0	2.8
P																		3.1	2.9
A																			2.7

Table A.3 – Side chain square well widths in Å (table 2.4, right side)

	L	I	V	M	F	Y	W	H	S	T	N	Q	D	E	K	R	C	P	A
L	6.4	6.5	6.2	6.5	6.6	6.7	6.9	6.5	6.3	6.2	6.4	6.3	6.5	6.4	6.5	6.8	6.1	6.3	5.6
I		6.6	6.4	6.7	6.6	6.8	6.8	6.6	6.4	6.4	6.6	6.6	6.5	6.6	6.7	6.7	6.4	6.4	5.7
V			6.3	6.4	6.5	6.5	6.6	6.2	6.2	6.4	6.3	6.5	6.3	6.5	6.6	6.8	6.0	6.3	6.1
M				6.7	6.5	6.6	7.0	6.5	6.4	6.4	6.4	6.4	6.7	6.4	6.4	6.6	6.3	6.2	5.8
F					6.8	6.8	7.0	6.5	6.2	6.6	6.5	6.6	6.7	6.8	6.9	6.9	6.4	6.5	5.9
Y						7.0	7.0	6.9	6.5	6.4	6.7	6.7	6.9	6.8	6.7	7.0	6.5	6.4	5.7
W							7.4	7.1	6.3	6.5	6.9	6.7	6.9	6.9	6.5	6.9	6.4	6.3	5.5
H								6.7	6.3	6.3	6.5	6.6	6.6	6.4	6.6	6.9	6.2	6.3	5.5
S									6.4	6.0	6.2	6.0	6.1	6.0	6.1	6.3	6.3	6.1	5.9
T										6.5	6.3	6.4	6.2	6.4	6.5	6.8	6.1	6.6	6.2
N											6.3	6.4	6.5	6.4	6.5	6.6	6.2	6.2	5.6
Q												6.6	6.3	6.6	6.7	6.9	6.1	6.5	5.8
D													6.5	6.6	6.3	6.5	6.2	6.3	5.6
E														6.7	6.4	6.6	6.1	6.4	5.9
K															6.9	6.8	6.4	6.7	6.0
R																7.2	6.3	6.8	6.1
C																	6.2	6.0	5.9
P																		6.5	6.2
A																			5.4

Table A.4 – Side chain bond and pseudobond lengths (in Å) and bead masses (relative to CH_3) (table 2.5)

	C_α–R	NH–R	CO–R	Mass
L	2.625	3.290	3.500	3.799
I	2.400	3.050	3.300	3.799
V	2.002	2.775	2.959	2.866
M	3.400	3.800	4.050	4.998
F	3.425	3.650	4.050	6.061
Y	3.843	4.050	4.300	7.126
W	3.881	4.100	4.350	8.66
H	3.160	3.450	3.830	5.394
S	1.967	2.650	2.800	2.064
T	1.981	2.650	2.900	2.997
N	2.510	3.050	3.350	3.862
Q	3.300	3.750	4.000	4.795
D	2.500	3.100	3.250	3.860
E	3.180	3.780	3.930	4.793
K	3.550	4.050	4.250	4.865
R	4.200	4.500	4.800	6.728
C	2.350	2.800	3.100	3.13
P	1.926	1.851	2.995	2.80
A	1.600	2.500	2.560	1.000

Table A.5 – Side chain squeeze diameters *sqz*6-10 in Å (table 2.7)

	*sqz*6 CO_{i-1}–R_i	*sqz*7 NH_{i+1}–R_i	*sqz*8 $C_{\alpha i+1}$–R_i	*sqz*9 $C_{\alpha i-1}$–R_i	*sqz*10 CO_{i-2}–R_i
L	3.918	3.724	4.863	4.936	5.001
I	3.740	3.626	4.867	4.867	4.994
V	3.570	3.378	4.635	4.754	5.002
M	4.205	4.032	5.067	5.206	5.017
F	3.991	3.973	4.827	4.780	5.040
Y	4.208	4.246	4.978	4.898	5.042
W	4.460	4.187	4.963	5.180	4.986
H	3.886	3.838	4.790	4.766	4.945
S	3.331	3.128	4.380	4.507	4.944
T	3.447	3.290	4.573	4.617	5.007
N	3.607	3.565	4.680	4.633	4.791
Q	4.139	3.996	5.062	5.134	5.000
D	3.751	3.435	4.558	4.785	4.860
E	4.175	3.997	5.074	5.162	4.996
K	4.384	4.191	5.163	5.323	4.974
R	4.827	4.651	5.535	5.703	4.978
C	3.516	3.350	4.501	4.560	4.913
P	3.133	3.298	4.665	3.884	4.773
A	3.312	3.000	4.353	4.598	4.997

Table A.6 – Side chain squeeze factors *sqz*6-10 and "original" bead diameters obtained by inverse application of the Lorentz-Berthelot combining rules (table 2.8).

	*sqz*6 CO_{i-1}–R_i	*sqz*7 NH_{i+1}–R_i	*sqz*8 $C_{\alpha i+1}$–R_i	*sqz*9 $C_{\alpha i-1}$–R_i	*sqz*10 CO_{i-2}–R_i
sqz	0.7607	0.7930	1.0956	1.1244	0.9259
L	6.301	6.092	5.177	5.080	6.802
I	5.833	5.845	5.185	4.957	6.787
V	5.386	5.220	4.761	4.756	6.805
M	7.056	6.869	5.550	5.560	6.837
F	6.493	6.720	5.112	4.802	6.887
Y	7.063	7.409	5.387	5.012	6.891
W	7.726	7.260	5.360	5.514	6.770
H	6.217	6.380	5.044	4.777	6.681
S	4.758	4.589	4.296	4.317	6.679
T	5.063	4.998	4.648	4.512	6.815
N	5.483	5.691	4.843	4.541	6.349
Q	6.882	6.778	5.541	5.432	6.800
D	5.862	5.363	4.621	4.811	6.498
E	6.977	6.781	5.563	5.482	6.792
K	7.526	7.270	5.725	5.768	6.744
R	8.691	8.430	6.404	6.444	6.753
C	5.244	5.149	4.517	4.411	6.612
P	4.237	5.018	4.816	3.209	6.310
A	4.708	4.266	4.246	4.479	6.794

Appendix B

Supplementary figures

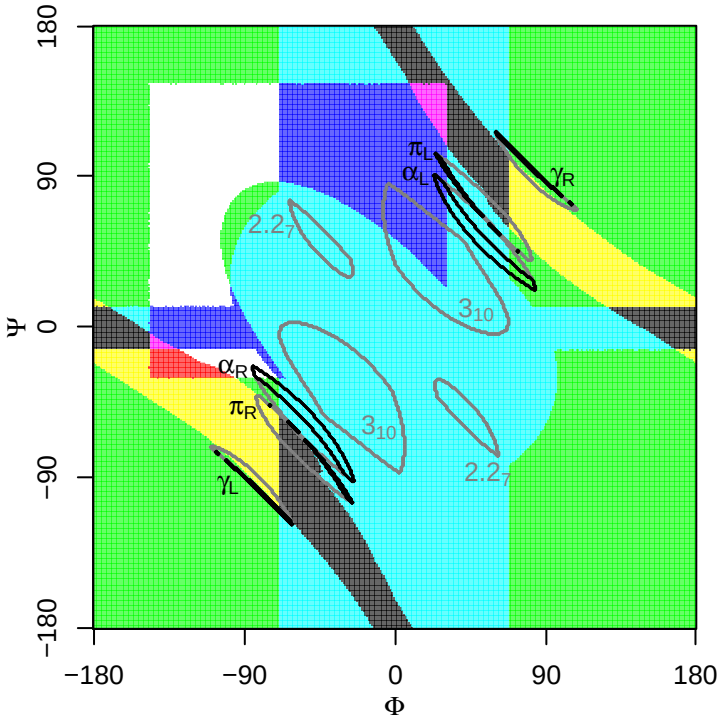

Figure B.1 – Ramachandran plot of polyalanine, following the colour scheme of figs. 5.5 and 5.8. A small section of the α_R-helical region is free, enabling polyA to form such helices. Unlike polyserine and polyglutamine, polyalanine cannot reach $\Phi = 180°$, hindering extended configurations and reducing the radius of gyration and end-to-end distance in the random coil state significantly compared to polyglutamine and slightly compared to polyserine (fig. B.4).

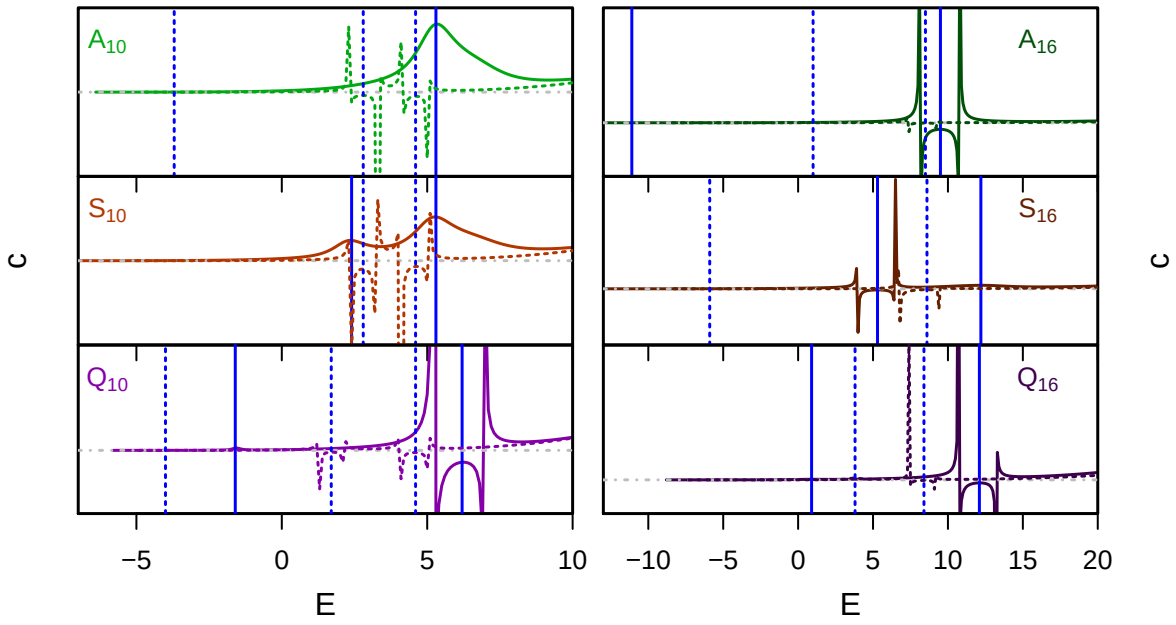

Figure B.2 – Microcanonical heat capacities of all simulated systems in the PRIME20n (continuous lines) and PRIME20s (dashed) models. The transition energies are marked by blue lines. The $N = 16$ transition temperatures are noted over the course of section 5.4. For $N = 10$ they are 0.152 (A_{10}), $\{0.139, 0.150\}$ (S_{10}), $\{0.064, 0.147\}$ (Q_{10}) in PRIME20n and $\{0.018, 0.105, 0.103\}$ (A_{10}), $\{0.110, 0.109\}$ (S_{10}), $\{0.017, 0.093, 0.105\}$ (Q_{10}) in PRIME20s. Even though S_{10} and A_{10} have two first-order signatures each, these signatures lie at almost identical temperatures (the ones with higher E even at lower T) and are not distinguishable physically.

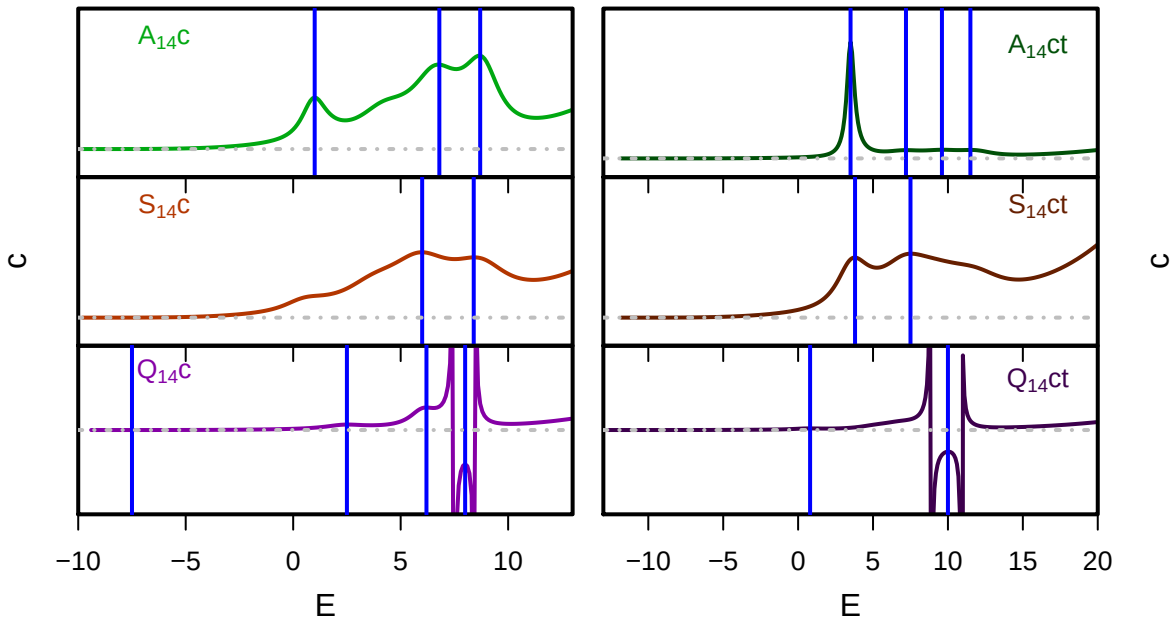

Figure B.3 – Microcanonical heat capacities of the $N = 14$ systems with chromophores and solubility tail. The corresponding X_{16} graphs are found in figures 4.4 and 4.12. For both polyA and polyS, the complex landscape of second-order transitions is largely unaffected by the modifications, but the first-order helix folding transition is replaced by a second-order signature. The polyQ behaviour is similar between all three systems, dominated by one hairpin folding transition.

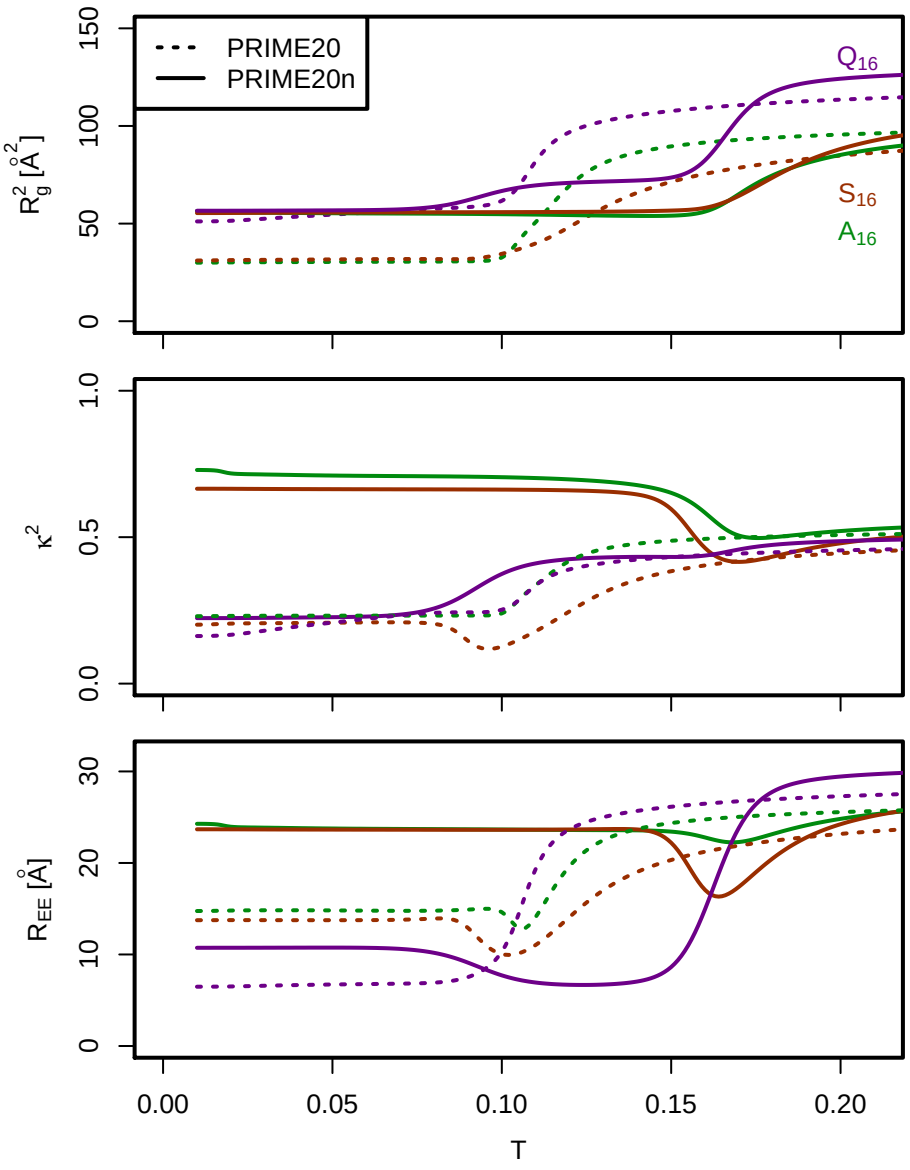

Figure B.4 – Squared radius of gyration $R_g^2(T)$, shape anisotropy $\kappa^2(T)$ and end-to-end distance $R_{EE}(T)$ of all $N = 16$ chains in PRIME20 (dashed lines) and PRIME20n. This figure sketches how the three quantities can be used to characterise the different states. $R_g^2(T)$ of A_{16} and S_{16} was also shown in fig. 4.7 and $R_{EE}(T)$ of Q_{16} and S_{16} in fig. 4.13. The α-helix state (S/A_{16} at low T in PRIME20n) is well recognisable due to characteristic values of R_{EE} and κ^2. The γ-helix (S/A_{16} at low T in PRIME20) on the other hand cannot be distinguished from the polyQ hairpins by means of κ^2, but it has a distinct R_g^2 value and also differs in R_{EE}. The globule states of S/A_{16} are practically invisible in R_g^2, but recognisable by a local minimum in R_{EE}. Overall, none of the three quantities is sufficient to achieve a complete picture of the behaviour of these systems, but all of them can be of use depending on the context.

Acknowledgements

I would like to thank everyone who made it possible for me to write this thesis and who helped in some way to complete it.

First of all, I am of course deeply grateful to my advisor, Prof. Wolfgang Paul, for his guidance throughout the past five years. He introduced me to an interesting topic to work on and always had a vision where the project was going to lead – which I was certainly lacking for most of the time. His availability to give competent answers and advice at practically any time and in any situation are not to be taken for granted (as I learned in discussions with PhD students from other groups and universities), and I am glad to have worked with him.

Next on the list is my mentor, Prof. Thomas Kiefhaber. As an experimental biochemist, he was often able to provide a new point of view for us theoretical physicists, and he too was available for discussion at the strangest of times – like mere hours after returning from a conference somewhere far away (India, if I remember correctly). In addition to all of this, he played an important role as the organiser of the annual "Faltertage" conference, putting our collaborative polyglutamine project on display to the international protein science community.

Within the Paul group, Dr. Benno Werlich was my primary source of information and technical advice. Without Benno, I would have had no idea how to begin the project or how to continue it at many points. Aside from this professional role, Benno has also been a major influence on my personal life, be it as the biggest bicycle nerd I know, as a music producer and radio host, or simply as an overall social person inside and outside of our office.

Another very important member of the group was and still is Dr. Timur Shakirov, the personification of all knowledge about SAMC, source of countless helpful ideas and humorous comments, and a valuable partner in the preparation and presentation of shared posters.

Paul Käthner began the simulations of polyserine during his Bachelor's thesis and, following Timur's discovery of the conversion from the configurational to the full microcanonical ensemble, went much deeper into this topic than I had done myself. Despite the limitations of a Bachelor's thesis, Paul's work has been immensely helpful for my own progress as well.

While Benno, Timur and Paul were my scientifically closest partners within the group, I would also like to thank the remaining current and former group members for their scientific and private support and discussions. Prof. Steffen Trimper, Prof. Viktor Ivanov, Dr. Semjon Stepanov, Dr. Eunsang Lee, Dr. Jeanette Köppe, Dr. Anja Kuhnhold, Michael Beyer, Lama Tannoury, Christian Lauer and Yara all had their influence on my work and life for which I am grateful. And even if he is not directly part of the group, I would like to include Dr. Jan Kantelhardt (another highly inspiring person) in this list as well.

Among these, Benno and Micha deserve another special mention for taking the time to proofread the dissertation. Thank you for all your questions and corrections!

Over the years, our group hosted many interesting international guests. Among these guests, Prof. Jutta Luettmer-Strathmann stands out because, having recently begun to work with PRIME20 as well, we had a very fruitful (and weirdly entertaining) discussion about the quirks and specialities of this model, which gave chapter 5 and the entire thesis a new direction.

I would like to thank Prof. Carol Hall, whose group designed PRIME20, who provided missing model parameters even in advance of their publication, and her recent graduate Dr. Yiming Wang, who kindly explained the squeeze factors to me while he was preparing for his own defence.

My project was funded by the DFG through the project SFB TRR-102. Although I do not want to underestimate the importance of enabling me to buy food, the contact to many other members of related groups in Halle and Leipzig is certainly the most valuable effect of this SFB. In this context, I especially would like to thank Dr. Thomas Michael, our PhD shepherd, for his dedication to lead his herd (called the iRTG) through all adversity towards successfully achieving the PhD.

If I tried to thank every iRTG member individually for being around, I would inevitably forget someone, so the personal mentions will be limited to Svetlana Pylaeva and Peter Enke, with both of whom I was happy to work together more closely. But of course everyone else added to the very nice, interesting and supportive community as well.

Finally, there are many people outside the scientific context who should be thanked – family, friends, colleagues... Above all, these are my parents and my sisters, who in so many ways paved the road towards the life I am privileged and thankful to live, and in the context of this work I want to end by thanking everyone else who was willing to listen to me palavering about computational biophysics.

Angaben zur Person

Name	Arne Böker
Geburtsdatum	30. November 1990
Geburtsort	Bielefeld
Geschlecht	männlich
Staatsangehörigkeit	deutsch
Adresse	Schleiermacherstraße 2a, 06114 Halle

Bildungsgang

10/2009 - 09/2012	Studium der Medizinischen Physik an der Martin-Luther-Universität Halle-Wittenberg. Abschluss: Bachelor of Science Thema der Bachelorarbeit: *Reconstruction of complex networks based on event time series*
10/2012 - 10/2014	Studium der Physik an der Martin-Luther-Universität Halle-Wittenberg. Abschluss: Master of Science Thema der Masterarbeit: *Wang-Landau simulations of protein-like Gō model molecules*

Publikationsliste

Arne Böker, Wolfgang Paul, *Wang-Landau simulation of Gō model molecules*.
Eur. Phys. J. E **39**(2016):5. doi: 10.1140/epje/i2016-16005-x

Timur Shakirov, Sergey Zablotskiy, Arne Böker, Viktor A. Ivanov, Wolfgang Paul, *Comparison of Boltzmann and Gibbs entropies for the analysis of single-chain phase transitions*.
Eur. Phys. J. ST **226**(2017):705. doi: 10.1140/epjst/e2016-60326-1

Svetlana Pylaeva, Arne Böker, Hossam Elgabarty, Wolfgang Paul, Daniel Sebastiani, *The conformational ensemble of polyglutamine-14 chains: specific influence of solubility tail and chromophores*.
ChemPhysChem **19**(2018):2931. doi: 10.1002/cphc.201800551

Arne Böker, Paul Käthner, Wolfgang Paul, *Thermodynamics and conformations of polyalanine, polyserine and polyglutamine. A comparison using the PRIME20 model*.
In preparation.

Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne fremde Hilfe verfasst, andere als die von mir angegebenen Quellen und Hilfsmittel nicht benutzt und die den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Halle, den 18. März 2019

Arne Böker