Understanding complex biomolecular systems through the synergy of molecular dynamics simulations, NMR spectroscopy and X-Ray crystallography

Tim Zeiske

Submitted in partial fulfillment of the

requirements for the degree

of Doctor of Philosophy

in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2016

ABSTRACT

Understanding complex biomolecular systems through the synergy of molecular dynamics simulations, NMR spectroscopy and X-Ray crystallography

Tim Zeiske

Proteins and DNA are essential to life as we know it and understanding their function is understanding their structure and dynamics. The importance of the latter is being appreciated more in recent years and has led to the development of novel interdisciplinary techniques and approaches to studying protein function. Three techniques to study protein structure and dynamics have been used and combined in different ways in the context of this thesis and have led to a better understanding of the three systems described herein.

X-ray crystallography is the oldest and still arguably most popular technique to study macromolecular structures. Nuclear magnetic resonance (NMR) spectroscopy is a not much younger technique that is a powerful tool not only to probe molecular structure but also dynamics. The last technique described herein are molecular dynamics (MD) simulations, which are only just growing out of their infancy. MD simulations are computer simulations of macromolecules based on structures solved by X-ray crystallography or NMR spectroscopy, that can give mechanistic insight into dynamic processes of macromolecules whose amplitudes can be estimated by the former two techniques.

MD simulations of the model protein GB3 (B3 immunoglobulin-binding domain of streptococcal protein G) were conducted to identify origins of discrepancies between order parameters derived from different sets of MD simulations and NMR relaxation experiments.

The results highlight the importance of time scales as well as sampling when comparing MD simulations to NMR experiments. Discrepancies are seen for unstructured regions like loops and termini and often correspond to nanosecond time scale transitions between conformational substates that are either over- or undersampled in simulation. Sampling biases can be somewhat remedied by running longer (microsecond time scale) simulations. However, some discrepancies persist over even very long trajectories. We show that these discrepancies can be due to the choice of the starting structure and more specifically even differences in protonation procedures. A test for convergence on the nanosecond time scale is shown to be able to correct for many of the observed discrepancies.

Next, MD simulations were used to predict *in vitro* thermostability of members of the bacterial Ribonuclease HI (RNase H) family of endonucleases. Thermodynamic stability is a central requirement for protein function and a goal of protein engineering is improvement of stability, particularly for applications in biotechnology. The temperature dependence of the generalized order parameter, S, for four RNase H homologs, from psychrotrophic, mesophilic and thermophilic organisms, is highly correlated with experimentally determined melting temperatures and with calculated free energies of folding at the midpoint temperature of the simulations. This study provides an approach for *in silico* mutational screens to improve thermostability of biologically and industrially relevant enzymes.

Lastly, we used a combination of X-ray crystallography, NMR spectroscopy and MD simulations to study specificity of the interaction between *Drosophila* Hox proteins and their DNA target sites. Hox proteins are transcription factors specifying segment identity during embryogenesis of bilaterian animals. The DNA binding homeodomains have been shown to confer specificity to the different Hox paralogs, while being very similar in sequence and structure. Our results underline earlier findings about the importance of the N-terminal arm and linker region of Hox homeodomains, the cofactor Exd, as well as DNA shape, for specificity. A comparison of predicted DNA shapes based on sequence alone with the shapes

observed for different DNA target sequences in four crystal structures when in complex with the *Drosophila* Hox protein AbdB and the cofactor Exd, shows that a combined "induced fit"/"conformational selection" mechanism is the most likely mechanism by which Hox homeodomains recognize DNA shape and achieve specificity.

The minor groove widths for all sequences is close to identical for all ternary complexes found in the different crystal structures, whereas predicted shapes vary between the different DNA sequences. The sequences that have shown higher affinity to AbdB *in vitro* have a predicted DNA shape that matches the observed DNA shape in the ternary complexes more closely than the sequences that show low *in vitro* affinity to AbdB. This strongly suggests that the AbdB-Exd complex selects DNA sequences with a higher propensity to adopt the final shape in their unbound form, leading to higher affinity.

An additional AbdB monomer binding site with a strongly preformed binding competent shape is observed for one of the oligomers in the reverse complement strand of one of the canonical (weak) Hox-Exd complex binding site. The shape preference seems strong enough for AbdB monomer binding to compete with AbdB-Exd dimer binding to that same oligomer, suggested by the presence of both binding modes in the same crystal. The monomer binding site is essentially able to compete with the dimer binding site, even though binding with the cofactor is not possible, because its shape is very close to the ideal shape.

A comparison of different crystal structures solved herein and in the literature as well as a set of molecular dynamics simulations was performed and led to insights about the importance of residues in the Hox N-terminal arm for the preference of certain Hox paralogs to certain DNA shapes. Taken together all these insights contribute to our understanding of Hox specificity in particular as well as protein-DNA interactions in general.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank many people who have contributed in some way to the this accomplishment, directly or indirectly. The quality of this work (and maybe even its existence) would be far inferior if it had not been for all of you.

First and foremost I want to acknowledge Dr. Arthur Palmer, my advisor over the last 6+ years. Not only is Art one of the smartest people I have ever known and a great inspiration, but he also gave me all the freedoms I could ask for while being an extremely supportive and hands-on mentor when I needed it. His insight and problem solving abilities are priceless when you are stuck with a problem and his patience and fairness seem unlimited.

Secondly, I want to thank my thesis committee, Drs. Wayne Hendrickson, Richard Mann, Ann McDermott and Lawrence Shapiro. Richard and Larry have been the most directly involved in my research and their advice and guidance has been tremendously helpful. Wayne and Ann are both brilliant minds and have always been helpful with any issues I have presented them with. Wayne is a limitless source of knowledge about X-ray crystallography (and many other things).

Other faculty have been involved in the research presented here but were not on the committee. The main ones I want to mention are Drs. Barry Honig, Richard Friesner and David Shaw who have contributed through both discussions and resources.

I would say that I was standing on the shoulders of giants during my time as a PhD student. But while this might be somewhat true with respect to the people mentioned above (who are indeed giants of their respective fields), much of my work was done standing on the shoulders of other grad students and post-docs.

In particular I need to thank Paul Harvilla and Jae-Hyun Cho, who are just much better at protein expression and purification than I will ever be. Then I need to thank Jae-Hyun Cho, Michelle Gill, Ying Li and Paul O'Brien (and of course Art), who are just much better at NMR than I will ever be. I also want to thank Paul Robustelli and especially Kate Stafford who are much better at MD simulations than I will ever be. If all of you are coauthors of this thesis (which you should be), then Kate is at least a second author, if not co-first (in particular for the two MD chapters in the thesis), given all the time she spent helping me with all kinds of things (oftentimes my own stupidity).

And finally the entire Shapiro (in particular Julia - without whose crystal picking skills I would have no structures -, Oliver, Gil, Filip, Kerry, Anna) and Sobolevsky labs who are all better at X-Ray crystallography than I will ever be (and of course Larry and Wayne, who spent a lot of time helping me with explanations about crystallography or actual hands on refinement advice).

I don't want to forget the amazing fly people in the Mann lab, in particular Namiko, Katie, Matt, Judith, Roumen and of course Nithya, who did much of the work in Chapter 6 (she solved two of the four structures).

And lastly, previous lab members that I barely knew, but whose theses were guiding me this whole time: Nikola Trbovic, Keri Siggers and Nichole O'Connell.

The list of people more indirectly involved is endless but I will give a try to honor some of them here: People in the Friesner and Honig labs and people at D.E. Shaw Research, the angels working at NYSBC (Mike, Kaushik, Shibani to name a few) who are always there to help if you have a problem at the NMR machines, and the amazing people here at CUMC who have made my life so much easier at many times (Fred Loweff, Ed Johnson, Rachel Hernandez, Aneudy Tapia, Wanda Noriega, Jessica Sama, Stacy Warren and others).

Now I would also like to thank my friends, who of course have helped in the creation of this thesis by keeping me sane. I will try to list some here but if you don't see your name here, don't be offended I will buy you a beer or something. My friends in New York are the ones that have helped most with my daily stress reduction. In particular: Adler, Alba, Amanda, Antonio, Ben, Cheryl, Fre (Carp & Frapido), Hojjat & Mersedeh (they always go together you cannot list them alphabetically), Jason, Matteo, Tiara, Titus (for a while), Rishi, Yiorgos. I am sure for each name here there are two I am forgetting. Thanks for working out with me, drinking with me, dancing with me, partying with me, discussing science, sports, technology, politics, philosophy with me, disagreeing with me, and just being around to keep me motivated.

Adler needs some special mention here not only because he is a fantastic friend (and mentor in many areas) but also because he took care of so many things for me while I was writing this thesis. Without him this thesis would probably not have happened (definitely not in time).

There are plenty of friends from other places I would like to mention but I will also have to leave it at just a few names here: Lenni, Paul, Leif, Tobi, Till, Julian, Krylle, Sascha, Felix, Boris, Nora, Titus and of course Robert Rosenberg. Special thanks to my Onzaner buddies Lenni, Leif, Tobi, Paul (and sometimes Jason, Julian, Krylle, Sascha and others as well) who I can always count on at least once a year to forget about all the *nonsense* (censored) in life and spend the day with what's really important: friends and beer.

Another person who needs a special mention is Zeynep. I didn't forget you :-). The list of things I need to be thankful for is longer than this thesis but here are a few: For very practical help in taking care of a lot of things for me during the last couple of weeks/months, and the entire PhD: Helping me find a suit, helping find an accommodation for my family, helping with food, pictures and other things at the actual defense, for organizing the party and other things that would have caused me stress because I wouldn't have been able to focus as much on writing as I should have. For emotional support and your constant assurances that everything would be fine and in general just being there for me when I was extremely stressed and felt in above my head. I am pretty sure I could not have finished this thesis without you!

Last but most definitely not least I want to thank my family. Without my family I would most definitely not be where I am today, and that includes this thesis, as well as many other things. I will also include Paul Barrett here because he is basically part of the family and has been great help with my applications to grad school, both by lending me his credit card for the application fees and reviewing my English. I am also including all the animals that have been part of this zoo of a household over the years, most of all Spike who I miss dearly to this day.

Der größte Dank gilt meiner Familie. Danke Lino, du bist der beste Bruder der Welt! Den größten Verdienst an dieser Arbeit haben ohne Zweifel meine Eltern. Ohne euch gäb es mich nicht, und deshalb auch nicht diese Dissertation. Ich danke euch von ganzem Herzen für eine schöne Kindheit und tolle Jugend, trotz aller Holprigkeiten. Papa, dir habe ich sicherlich mein Interesse an der Wissenschaft zu verdanken. Unsere stundenlangen Gespräche über Gott und die Welt, Atome und Galaxien, haben mich schon früh gelehrt alles zu hinterfragen, jeden Stein umzudrehen und zu probieren, alles zu verstehen. Und zu guter Letzt danke Mama, du warst wirklich immer für mich da und hast mich in allem unterstützt, egal wie dumm es (ich) war. Ob ich eine Rakete im Keller bauen wollte, alleine mit der U-bahn nach Tiergarten fahren wollte, alleine mit meinen Freunden in den Urlaub nach Spanien wollte oder meine Doktorarbeit in den Vereinigten Staaten machen wollte. Ich weiß, dass es (ich) nicht immer einfach war, aber ich hoffe du siehst, es hat sich gelohnt.

Chapter 1

Background

Proteins are an integral part of living cells, carrying out a vast array of functions. While all proteins have a dynamical character, some proteins carry out their designated functions based mainly on their three dimensional structure alone. Examples of such proteins are proteins of the cytoskeleton or the nucleosome. Many if not most proteins, however, can function only because of their dynamical character, meaning they function by changing their three dimensional structure on different time scales. Proteins in this latter group include enzymes, receptors, motor proteins and many more. To understand how proteins work at a molecular level, we thus have to not only look at their static structure but also at their dynamic properties. No single method in the field of structural biology has all the answers, which is why structural biologists rely more and more on a combination of different methods to study the functions of proteins (and other macromolecules). Three methodologies will be described here, namely X-ray crystallography, NMR spectroscopy and Molecular Dynamics (MD) simulations. All three cover a large area of aspects of protein structure and function, with many overlaps, forming a synergistic Venn diagram that is a powerful tool in understanding protein function at a molecular level.

1.1 X-ray crystallography

X-ray crystallography is the oldest and still most popular method to determine the structure of macromolecules and in particular proteins. The first protein structure determined by X-ray crystallography was that of sperm whale myoglobin in 1958 at a resolution of 6Å [1]. The progress made since then becomes apparent when comparing this early structure to the atomic resolution (1Å) structure of the same protein published in 1999 [2]. This improvement is not only due to improved equipment (in particular detector technology and the use of synchrotrons as light sources) but also the development of powerful algorithms and computer programs, together with ever more powerful computers.

In a classical single crystal X-ray diffraction experiment, a monochromatic beam of X-rays irradiates a protein (or other molecule) crystal. While most of the beam will fail to interact with the crystal and continue in a straight path, a fraction of the X-ray interacts with the matter in the crystal and is scattered into different directions. The scattered beams can be recorded on film or on an electronic detector as so-called diffraction spots (or ”reflections”) forming a diffraction pattern. Even though X-rays are chosen to have a maximal fraction of the light interact with the molecules in the crystal (a typical wavelength used will be around 1Å, which is on the scale of the interatomic distances), the diffracted light of a single molecule would be much too weak to detect. This cannot be counteracted by irradiating the same molecule for a long time, because the interaction of the beam with the molecule will slowly destroy the molecule. This is why crystals are used. Crystals by definition are one-, two- or three-dimensional lattices with many repeats of the same structure (called the unit cell) in all directions. Taking the crystals in the present study as an example, which were far from the biggest available protein crystals with dimensions of maybe 10 x 50 x 200 μm and unit cell edges of 50 - 100 Å, we can estimate the number of unit cells to be about 4x10^{11}, with one or two protein-DNA complexes per unit cell, yielding a total of almost a trillion molecules of interest per crystal.

While most of the diffracted electromagnetic waves interfere destructively and will still not result in measurable signal, the arrangement of the molecules in the crystal is not random, but a certain structural motif is repeated over and over as translational copies in a three dimensional lattice, resulting in positive interference of diffracted beams in a few specific directions defined by the type of lattice in the crystal. The directions in which interference is constructive (and thus locations of the diffraction spots) will be determined solely by the crystal lattice (and thus the geometry of the unit cell) and the wavelength of the X-rays, while the effectiveness of the constructive interference (and thus the intensity of the diffraction spots) is determined by arrangement of the atoms inside of the unit cell. By knowing the wavelength of the X-rays we can thus determine the dimensions of the unit cell by analyzing the locations of the reflections. Using the information about the unit cell together with the intensities of the measured reflections we can infer the arrangement of the atoms inside of the unit cell (*vide infra*).

The most intuitive way to understand the relationship of the diffraction pattern with the crystal lattice is Bragg's law (Figure 1.1 and equation 1.1). Bragg's law simply states, that because the scattering event is elastic (the wavelength of the X-rays is not changed by the event) interference of reflected light waves from two successive planes of the crystal lattice is constructive if and only if the path difference of the two light beams is a multiple of the used wavelength. Because the angle of incident and reflected light waves with the planes of the crystal are identical, simple trigonometry can be used to formulate Bragg's law as follows:

$$n\lambda = 2dsin\theta \tag{1.1}$$

where n is any natural number, λ is the wavelength of the light beam, d is the distance between the two planes that are reflecting the two interfering light waves.

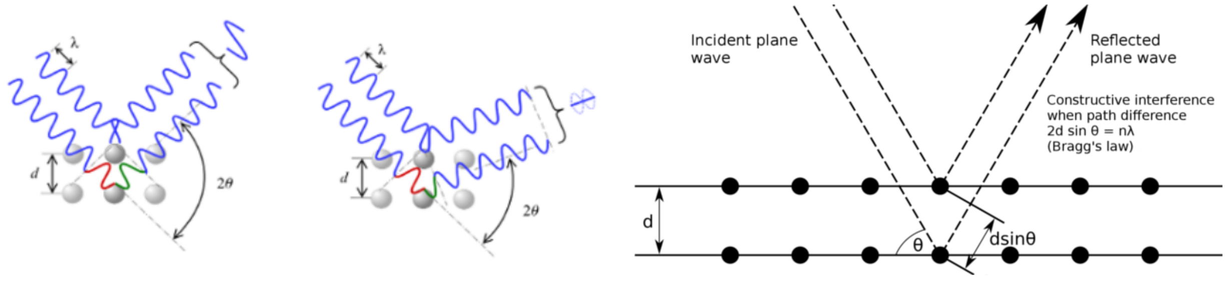

λ
d
2θ
λ
d
2θ
Incident plane wave
Reflected plane wave
Constructive interference when path difference 2d sin θ = nλ (Bragg's law)
d
θ
dsinθ

Figure 1.1: **Bragg's law**
An illustration of Bragg's law adapted from [3]. Elastic scattering on atoms of parallel planes of the crystal lattice. Elastic scattering means that the energy and thus wavelength of the light is identical before and after the scattering event. Because the wavelength does not change, and because the angle of the incident and the reflected waves to to the planes are identical, we can formulate the criterion for constructive interference by requiring the path difference of the reflected waves of two successive planes to be a multiple of the used wavelength. Using simple trigonometry we can state Bragg's law as follows: $n\lambda = 2dsin\theta$, where n is any natural number, λ is the wavelength of the X-ray beam, d is the distance between the planes and θ is the angle of the incident (or reflected) wave with the crystal plane.

Bragg's law is an intuitive way to understand the condition of constructive interference, but is not entirely accurate for the diffraction on a three-dimensional crystal, nor is it practical to solve a crystal structure. In particular, every atom that interacts with the X-rays (through its electrons) will scatter waves in all three dimensions. A more general version of Bragg's law are the von Laue conditions, given by:

$$\mathbf{a} \cdot (\mathbf{k} - \mathbf{k}') = 2\pi h \tag{1.2}$$

$$\mathbf{b} \cdot (\mathbf{k} - \mathbf{k}') = 2\pi k \tag{1.3}$$

$$\mathbf{c} \cdot (\mathbf{k} - \mathbf{k}') = 2\pi l \tag{1.4}$$

where **a**, **b** and **c** are the primitive vectors of the crystal lattice, **k** is the wave vector of the incident light beam and **k'** of the scattered beam. The resulting difference vector $\mathbf{Q} = (\mathbf{k} - \mathbf{k'})$ is also called the scattering vector and is perpendicular to the diffracting plane. h, k and l are integer numbers (also called Miller indices) and each (hkl) triplet corresponds to a specific reflection in the diffraction pattern where the von Laue conditions are met.

We now know how the crystal lattice relates to the reflections on the detector, but not how the intensity of the reflections relates to the contents of the unit cell. If we consider an X-ray waves that is being scattered by a volume element of the unit cell, the path difference α of the scattered light wave to a wave scattered at the origin of the unit cell, is given by:

$$\alpha = 2\pi \frac{\hat{\mathbf{k}} - \hat{\mathbf{k}}'}{\lambda} \cdot \mathbf{r} = 2\pi \mathbf{S} \cdot \mathbf{r} \tag{1.5}$$

where **r** is the position vector of the volume element compared to the origin, $\hat{\mathbf{k}}$ is the normalized wave vector of the incident beam and $\hat{\mathbf{k}}$**'** the normalized wave vector of the reflected beam (both vectors have a length of 1). The resulting vector $\mathbf{S} = \frac{\hat{\mathbf{k}} - \hat{\mathbf{k}}'}{\lambda}$ is perpendicular to the reflecting plane of the crystal. Every possible vector **S** thus corresponds to a specific plane of the crystal lattice and a specific reflection within the diffraction pattern and a triplet (hkl) defined by the von Laue conditions.

If for a particular reflection, defined by a particular vector **S**, we integrate the wave functions of all scattered waves in that direction for all volume elements in the unit cell, we obtain the so called structure factor F, given by:

$$F(\mathbf{S}) = \int\limits_{\mathbf{r}} \rho(\mathbf{r}) \mathrm{e}^{2\pi i \mathbf{S} \cdot \mathbf{r}} d\mathbf{r} \tag{1.6}$$

where $\rho(\mathbf{r})$ is the electron density of each volume element d**r**. This means that the structure factors corresponding to the reflections are a Fourier transform of the electron density inside of the unit cell.

From this we can now compute the electron density as a function of the structure factors by calculating the inverse (discrete) fourier transform:

$$\rho(\mathbf{r}) = \frac{1}{V} \sum_{hkl} F(hkl) \mathrm{e}^{-2\pi i \mathbf{S} \cdot \mathbf{r}} \tag{1.7}$$

This means that for each volume element in the unit cell all reflections have to be taken into account. That is why in X-ray crystallography we can never solve only a part of the unit cell by considering only a part of the reflections. The Fourier transform makes it such that all atoms in the unit cell influence the intensity of each reflection and the other way around.

Another problem we face now is that the structure factors F(hkl) are complex numbers, but we can only measure the intensities I(hkl) = $|F(hkl)|^2$ of the reflections. That means that we know about the amplitude of the electromagnetic wave hitting the detector but not its phase. This is the so called phase problem of X-ray crystallography. Many solutions for the phase problem exist, including isomorphous replacement and anomalous dispersion experiments, which will not be explained here. The method of choice for many problems, as well as the one used in this study, is molecular replacement.

Molecular replacement uses the so called Patterson function, which is the Fourier transform of the intensities instead of the structure factors, which also turns out to be the convolution of the electron density with its inverse. The peaks in the Patterson function are the interatomic distances of the unit cell. If we know the three-dimensional structure of a molecule that is closely related to the molecule we are trying to solve, given the correct orientation of the molecule in the unit cell, their Patterson functions should be strongly correlated. For molecular replacement we can thus reduce the problem to six dimensional search of the position of a related molecule inside of our unit cell, using three rotation and three translation functions and minimizing the differences between the Patterson map of our diffraction data and the map derived from the model.

The result of molecular replacement is a model of our unit cell, that is related to the actual unit cell in the crystal but not identical. Further steps will try to make the model as similar as possible to the actual unit cell by rebuilding parts of the model manually and improve the overlap of structure factors resulting from the model with the recorded intensities. This process is called refinement.

As a criterion for how well the model fits the data, the so called R value is used, which is defined as follows:

$$R = \frac{\sum ||F_{obs}| - |F_{calc}||}{\sum |F_{obs}|} \tag{1.8}$$

where $|F_{obs}|$ are the magnitudes of the observed structure factors (square roots of the intensities) and $|F_{calc}|$ the magnitudes of the structure factors calculated from the current model. For an ideal model the R value would thus be 0, while for a random fit the R value will be around 60% (0.6). Any manual changes to the model during refinement will not only take into account knowledge about what makes chemical and biological sense but also try to minimize the R value. To avoid artificially overfitting the model to the data, usually a small percentage of the reflections (5-10%) are removed at random from the dataset and excluded from refinement. The 90-95% of reflections are then used for the so called R$_{work}$ and the randomly selected 5-10% for a so called R$_{free}$, which is used for an independent objective assessment of the quality of the model.

A few important steps of the data processing have been skipped until now. The first step is called indexing and corresponds to the assignment of the reflections to their respective planes in the crystal and the determination of the type of crystal lattice and the unit cell dimensions. The von Laue conditions (equations 1.2 - 1.4) can be used to construct a hypothetical "reciprocal" lattice of vectors $\mathbf{a}^*$, $\mathbf{b}^*$ and $\mathbf{c}^*$, which are perpendicular to the planes formed by the unit cell vectors in ”real” space (that is the actual crystal lattice) and have magnitudes that are inverse to the magnitudes of the unit cell vectors in real space. This model is useful because intersections of the reciprocal lattice with the so called ”Ewald sphere” correspond exactly to the positions of the reflections that are observed on the detector. The Ewald sphere is a theoretical sphere with a radius of $1/\lambda$ whose origin sits along the path of the incident beam and whose shell intersects with the origin of the reciprocal lattice. As the crystal is rotated in the X-ray beam, we simultaneously rotate the reciprocal lattice and different planes of it will intersect the Ewald sphere at different crystal orientations angles causing the diffraction patterns recorded by the detector. By analyzing the positions of the reflections we can thus calculate the reciprocal lattice using the wavelength of the X-rays and the distance of the detector to the crystal and thus infer the lattice dimensions in real space and the dimensions of the unit cell. This construction directly results from the von Laue conditions and Bragg’s law. A side product of the indexing process is that it will not only tell us about the unit cell dimensions but also about the ”space group” of the crystal, which tells us about symmetries within the unit cell by observing symmetries in the diffraction patterns. There are 65 allowed space group for protein crystals.

The data is then integrated, which means the different images at different orientations of the crystal are combined to a single dataset containing a list of reflections with their hkl indices and their intensities. Because reflections are not infinitely thin, some reflections will appear on several images as we rotate the crystal and the reciprocal lattice cuts through the Ewald sphere. This is because neither the X-rays nor the crystal are perfect and the intersections of the Ewald sphere and the reciprocal lattice will have a certain ”width” to them. In addition, for all crystals that have additional symmetries within the unit cell (meaning all space groups except most simple one, P1, which has no additional symmetries inside of the unit cell) will have symmetry equivalent reflections. Recognizing multiple observations of the same reflections and combining them into a single reflection is called merging. At the same time the intensities of the merged reflections are adjusted in a process called scaling.

As multiple observations of reflections are merged and scaled, the quality of the process is assessed by another kind of R value called R$_{sym}$ (sometimes R$_{merge}$), which is defined similarly to the R value defined in equation 1.8:

$$R_{sym} = \frac{\sum_{hkl} \sum_{j} |I_{hkl,j} - \langle I_{hkl} \rangle|}{\sum_{hkl} \sum_{j} I_{hkl,j}} \tag{1.9}$$

1.2 Nuclear magnetic resonance spectroscopy

While Nuclear Magnetic Resonance (NMR) spectroscopy can be used to solve three dimensional structures of proteins [4], its real power lies in its sensitivity to dynamic processes on many time scales [5–8] (Figure 1.2). Because there are a large number of NMR experiments for a large number of different applications, we will only use this chapter to describe a few basic principles needed to understand the experiments that were used in this dissertation.

NMR relies on the interaction of NMR active nuclei with the magnetic component of radiofrequency radiation. NMR active nuclei are nuclei that have a non zero spin number I, which depends on the number of protons and neutrons in the nucleus. Nuclei with a non zero spin number will have a net magnetic dipole and be able to interact with an external magnetic field. Inside of a static magnetic field, the net magnetic dipole will adopt a quantitized number of possible orientations. The number of possible orientations is given by 2I+1, where I is the spin number, which can have have values of any half or whole integer number (or zero). Each orientation is given a magnetic quantum number m, which corresponds to the potential energy of that particular state in an external magnetic field:

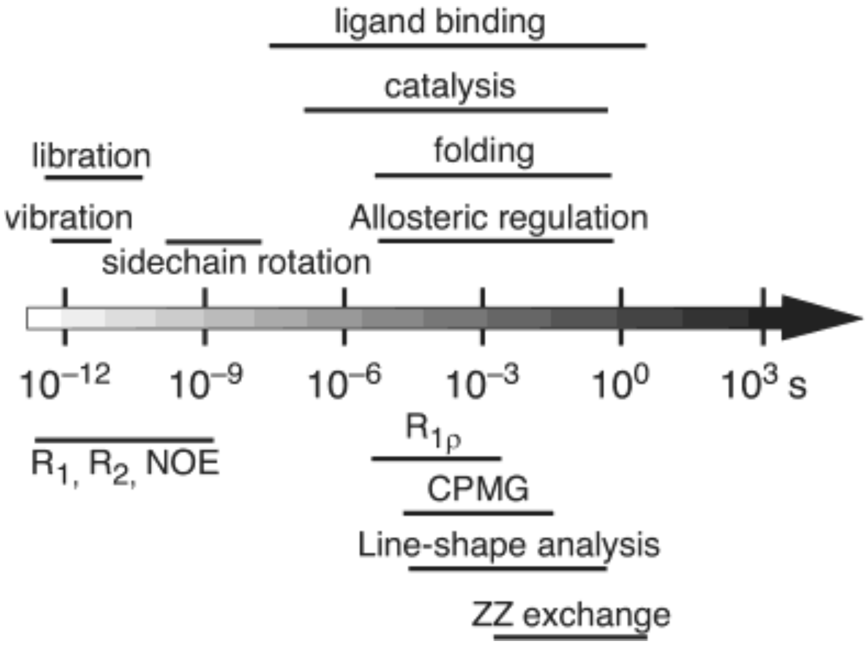


Figure 1.2: **Time scales of molecular motions accessible to NMR spectroscopy** Time scales of molecular motions observed in biomacromolecules, their corresponding functional relevance (top) and NMR techniques sensitive to them (bottom). Reprinted with permission from Chemical Reviews 104: 3623 - 3640. Palmer AG (2004) NMR characterization of the dynamics of biomacromolecules [8]. Copyright 2004 American Chemical Society.

$$E = -\boldsymbol{\mu} \cdot \boldsymbol{B} \tag{1.10}$$

where E is the potential energy of the magnetic dipole, $\boldsymbol{\mu}$ the magnetic dipole vector and $\boldsymbol{B}$ the magnetic field vector of the external field. Placing the external field along the z-axis of the laboratory frame we can calculate the z-component of the magnetic dipole interacting with the field as follows:

$$E = -\mu_z B_0 = -\gamma \hbar m B_0 \tag{1.11}$$

where μ_z is the z-component of the magnetic dipole, B_0 is the field strength of the external magnetic field (defined to lie along the z-axis), γ the so called gyromagnetic ratio (specific to each nucleus type), and m the magnetic quantum number of the particular state of the magnetic dipole (its quantitized orientation in the external field).

The different states (orientations) of the dipoles in solution are populated according to the Boltzmann distribution:

$$\frac{N_a}{N_b} = -\exp(-\frac{\Delta E}{kT}) \tag{1.12}$$

where N_a and N_b are the population of two spin states, ΔE the difference in energy between the two states, k the Boltzmann constant and T the temperature of the sample. Because the energy difference is given by $\Delta E = \gamma\hbar(m_a - m_b)B_0$, the population difference and thus the sensitivity of the NMR experiment increases with the strength of the external magnetic field.

In a typical NMR experiment, we do not measure individual nuclei but the so called bulk magnetization of the sample, which is the net magnetic dipole of the nuclei of interest in the sample. At equilibrium, the net magnetization **M** will be aligned with the external magnetic field. If we apply a second magnetic field, typically a short pulse of radiofrequency radiation in the xy-plane (rf pulse), this will cause the net magnetization vector to tilt away from the z-axis. The new vector **M**, which is not aligned with the z-axis anymore, will now precess about the external magnetic field vector (along the z-axis) at a frequency ω_0 defined by the gyromagnetic ratio of the nucleus of interest:

$$\omega_0 = -\gamma B_0 \tag{1.13}$$

The frequency ω_0 is specific to each nucleus type as defined by its gyromagnetic ratio, and is called its Larmor frequency.

The perturbation of the system through the rf pulse results in a non Boltzmann distribution of the spin populations. After some time, the net magnetization will return to the z-axis and thus the populations to Boltzmann equilibrium, through a process called spin-lattice relaxation (also "longitudinal relaxation"). The rate of this process is called the spin-lattice relaxation constant and is often denoted R_1 ($R_1 = 1/T_1$, with T_1 being the spin-lattice relaxation time). Spin-lattice relaxation involves the exchange of energy of the perturbed magnetization vector with the surroundings (the lattice). Spin-lattice relaxation occurs when molecular motions create time-dependent magnetic fields with contain a component at the eigenfrequencies of the spin system.

What is usually measured in an NMR experiment is the precession of the magnetic field around the static field. This precession induces an electrical current in a coil positioned perpendicularly to the z-axis of the spectrometer (on the xy-plane) and can thus be quantified. The measured signal is this an oscillating electrical current that gets weaker as time goes by and Boltzmann equilibrium is restored, i.e. a dampened sine wave. Another relaxation process contributes to the dampening of the signal, namely a process called spin-spin relaxation (also "transverse relaxation"), characterized by its relaxation constant R_2 ($R_2 = 1/T_2$), which usually dominates R_1 in protein NMR experiments. This process does not involve exchange of energy with the surroundings but is due to the dephasing of the bulk magnetization in the xy-plane. This dephasing is due to random time-dependent local fluctuations in magnetic fields experienced by each spin in the sample causing a loss of coherence between the different spins over time, because they precess at slightly different rates, resulting in a net loss of precessing magnetization, which can be measured by the coil in the xy-plane.

One of the important concepts of NMR spectroscopy is the so called chemical shift. The chemical shift is due to differences in the chemical environment of each spin (differences that are constant and not random and time-dependent as for spin-spin relaxation) leading to slightly different precession frequencies of the different types of spins. Protons that are part of aromatic rings for example will have substantially different local magnetic environments than protons that are part of aliphatic chains, leading to different precession frequencies and thus chemical shifts. The influence of local charge distributions on local magnetic fields is also called ”chemical shielding” and is defined as an offset to the Larmor frequency:

$$\omega = -(1 - \sigma)\gamma B_0 \tag{1.14}$$

where σ is the chemical shielding of a particular nucleus, which accounts for local differences in the magnetic field.

Because the resonance frequencies of the different nuclei are dependent on the strength of the external magnetic field, we usually use a reference compound with many equivalent protons that resonate at very high frequencies (tetramehthylsilane - TMS, or 2,2-methyl-2-silapentane-5-sulfonate - DSS, are frequently used ones), and define the resonance frequency of all other protons in reference to that compound, by the so called chemical shift, expressed in parts per million (ppm):

$$\delta = 10^6 \times \frac{\omega - \omega_{ref}}{\omega_{ref}} \tag{1.15}$$

where δ is the chemical shift, which is field independent.

Most modern NMR experiments are Fourier transform NMR experiments. This means, instead of continuously varying the rf field and measuring the resonance of all the different nuclei in the sample one by one (this is called a continuous wave experiment), a short pulse of rf radiation excites all the spins at once. The measured signal in the receiver coil is a linear combination of all the different precession frequencies. By Fourier transforming the signal from the time domain to the frequency domain, one obtains a full one-dimensional frequency domain spectrum (see chapter 5).

If this simple spectrum was recorded however, it would be dominated by the signal of water. Pure water has a molarity of 55.5 mol/l, which is many orders of magnitude above any solute in the sample. Many methods for water suppression exist. One of them is the jump return pulse sequence, which was used in chapter 5. The jump return experiment sets the carrier frequency on the solvent (water) resonance. After a 90° rf pulse that brings all spins into the xy-plane and a delay $\tau = 1/(4\Delta\nu_{max})$ a -90° pulse brings the water back to the z-axis, without affecting the spins of interest. $\Delta\nu_{max}$ is the difference in resonance frequency of the carrier (water resonance) and the signal of interest. The delay $1/(4\Delta\nu_{max})$ makes it such that the signal of interest has gained a phase difference of 90° in the xy-plane to the water magnetization, meaning that the second -90° pulse will bring the water magnetization back to the z-axis without affecting the signal of interest. This maximizes the signal of interest while minimizing the signal from water.

A perfect NMR spectrum would have infinitely narrow peaks at the positions of each type of nucleus (positioned at their respective chemical shifts). Because of the process of relaxation, and in particular spin-spin relaxation, the peaks have a finite width to them. This results from Fourier transforming a decaying sinusoidal signal. The linewidth is proportional to the transverse relaxation rate R_2 and linearly correlated with the size of the molecule in solution. This is a problem for protein NMR spectroscopy because it inherently limits the size of the proteins we can easily study by NMR.

As mentioned before, transverse relaxation is due to time-dependent random fluctuations in the local magnetic fields experienced by each spin. Different molecules in solution will be in different orientations compared to the external magnetic field, causing in particular dipolar interactions to vary for different equivalent spins in the sample, but also causing spins to experience different chemical shifts depending on their orientation with respect to the external field (”chemical shift anisotropy”). Due to the time-dependence and stochasticity of the involved processes, transverse relaxation is generally non recoverable.

The proportionality relationship of R_2 and the molecular size of the complex is a result of the slower tumbling of larger molecules. Small molecules will tumble relatively quickly, averaging out differences experienced by different spins. The bigger the protein and thus slower the tumbling of the protein in solution, the more this averaging process is inefficient, leading to broad lines in the spectrum. This is because tumbling needs to be fast compared to the processing leading to relaxation, in order for averaging to be efficient. This relationship allows us to estimate the size of a complex in solution, an idea that is used in chapter 5. By measuring the transverse relaxation time, we can estimate the size of a complex in solution and thus measure complex formation.

An experiment that is used to measure the transverse relaxation time is the Hahn echo experiment. In its simplest form, a 90° pulse brings the net magnetization to the xy-plane. During a first delay period τ phase differences between spins are accumulated both due to time independent phenomena such as chemical shielding or constant field inhomogeneities and random processes involved in transverse relaxation. Then a 180° pulse is applied, which inverts all magnetization. A second delay period τ, identical to the first, will allow phase differences from non random processes to rephase (because all of them were inverted by the 180° pulse), while transverse relaxation processes will continue to dephase the different spins. The signal is then recorded. The result of this experiment is that longer delay times will result in a reduced signal which is purely due to stochastic relaxation processes. By varying the delay one can thus estimate the relaxation time by fitting the signal intensities to an exponential decay function, thereby allowing us to estimate the tumbling time and molecular mass of the studied molecule or complex.

Because Fourier transform NMR techniques allow us to collect full spectra relatively quickly, it opens up the possibility for so called multi-dimensional NMR experiments, in which one or more variables of the experiment are varied in a linear fashion and treated as an additional dimension in the resulting spectrum. The variable that is usually varied is some time delay, yielding a multidimensional signal with time as a unit in all dimensions.

Multidimensional Fourier transformation then yields a multidimensional spectrum in the frequency domain for all dimensions.

One example is the NOESY experiment (Nuclear Overhauser effect spectroscopy), which in its simplest version is a three pulse experiment. A first 90° pulse applied to one particular spin I, brings the magnetization of this spin onto one of the transverse axes, for example the y-axis. This spin precesses at its specific precession frequency, which is dependent on its chemical shift. After a delay period t_1, another 90° pulse is applied along the same axis as before, which flips the y-component of the precessing spins back onto the z-axis. The important concept here is that the amount of the magnetization that lies along y when this second pulse is applied depends on both the time delay t_1 and the precession frequency, and thus the chemical shift, of the spin I. During a constant delay period τ_m following this second pulse, some of the magnetization of I that now lies along the z-axis, is transferred to another spin S, which is then flipped into the xy-plane by a 90° pulse, and its signal recorded. The process of the magnetization transfer here is the Nuclear Overhauser effect, which is based on dipolar interactions of spins through space and is strongly dependent on the internuclear distances and can thus be used to measure distances between nuclei.

This experiments correlates the precession frequency and thus chemical shift of the second spin S (which is measured during signal acquisition) to the chemical shift of the first spin I, which is measured indirectly by varying the delay t_1. A two-dimensional Fourier transform of the signal will thus result in a 2D spectrum with the frequencies of the two spin types along the two axes. Cross peaks off the diagonal of the spectrum correspond to nuclei that are in close proximity to each other. This method is used in chapter 5, to do a so-called NOESY walk through the sequence of a DNA oligomer, allowing us to assign all the peaks in a 2D imino spectrum of a DNA oligomer by knowing which nuclei lie close to each other.

A related idea is used in another very popular 2D experiment called HSQC (Heteronuclear single quantum coherence spectroscopy), which in its most popular form measures correlations between protons and nitrogens that share a chemical bond. In particular, because of the nature of proteins and peptide bonds, there is approximately one peak per amino acid in an HSQC spectrum, which is thus considered a fingerprint spectrum of a protein. Instead of through-space magnetization transfer, HSQC experiments rely on through bond magnetization transfer ("scalar coupling"), which allows correlation between nuclei that are attached to each other. In a similar manner as for the NOESY experiment, magnetization is transferred between a directly detected nucleus type (usually ^{1}H, termed "I") and an indirectly detected one (often ^{15}N, termed "S"). Magnetization is transferred from the I spin to the S spin with a pulse sequence called INEPT (Insensitive nuclei enhanced by polarization transfer), followed by a variable delay t_1, which contains an 180° pulse on the I spin, such that only the chemical shift of the S spin evolves during t_1. A reverse INEPT sequence is used to transfer magnetization back to the I spin, which is then detected. This experiment thus correlates the precession frequency of S spin which evolves during t_1 to the precession frequency of the I spin which evolves during acquisition of the signal. If this sequence is applied to the nuclei ^{1}H and ^{15}N, the resulting spectrum will have one peak per amino group in the sample, representing all amino acids except prolines, and a few side chains that have amino groups. The exact position of each peak corresponds to the local protein conformation. It can thus be seen why this spectrum is often referred to as the fingerprint of a protein.

Similar experiments are possible for other types of nuclei but shall not be discussed here. In particular we use a similar type of experiment called a methyl TROSY experiment which correlates ^{1}H nuclei to ^{13}C nuclei in methyl groups, following similar principles (chapter 5). In short, a TROSY experiment is designed in such a way that different processes contributing to transverse relaxation will cancel each other, leading to longer transverse relaxation times and thus sharper peaks in the spectrum.

1.3 Molecular dynamics simulations

As opposed to X-ray crystallography and NMR spectroscopy, molecular dynamics (MD) simulations are normally used to gain mechanistical insight into dynamical processes and not structural information. A crystal or NMR structure is usually needed to prepare an MD simulation. While NMR and even X-ray crystallography contain some information about dynamics, no other method has the temporal and spacial resolution of MD simulations. In addition, while NMR spectroscopy can provide information about the amplitudes of motions averaged over many molecules in solution, MD simulations provide directionality and thus causality of molecular motions for single molecules at atomic resolution. This makes MD simulations a powerful tool when used in synergy with experimental methods such as X-ray crystallography and NMR spectroscopy.

The first MD simulation of a protein was conducted in 1977 [9]. This first simulation of the bovine pancreatic trypsin inhibitor lasted for only 9.2 ps. Today MD simulations are routinely conducted for many micro- or even milliseconds, reaching protein folding time scales for fast folding proteins [10].

In its simplest form, an MD simulation takes a predefined set of atoms with defined interactions between them and integrates Newton's equations of motion to update each atom's position and velocity. This is repeated for many time steps as long as the researcher desires and his computational equipment allows. A simplified description of the process is as follows:

1. Set up a "box" containing all the desired molecules, for example an all atom representation of the protein of interest surrounded by many solute molecules (for example water molecules).

2. Define initial positions for all atoms at t = 0: $\vec{r}(t=0) = [\vec{r}_1, \vec{r}_2, ..., \vec{r}_N](t=0)$

3. Define initial velocities to all atoms by sampling randomly from a Boltzmann distri-

bution at the desired temperature: $\vec{v}(t=0) = [\vec{v}_1, \vec{v}_2, ..., \vec{v}_N](t=0)$

4. Define an interaction potential between all atoms of the system, based on "force field" (*vide infra*)

5. From the potential calculate the force experienced by each atom i: $\vec{F}_i = -\sum_{j=1}^{N} \nabla V_i(r_{ij}, i \neq j)$

6. Update the positions of all atoms by integrating the potential: $\vec{r}(t+\Delta t) = \vec{r}(t) + \vec{v}(t)\Delta t + \frac{1}{2}\vec{a}\Delta t^2 + ...$

7. Update the velocities of all atoms accordingly: $\vec{v}(t+\Delta t) = \vec{v}(t) + \vec{a}(t)\Delta t + ...$

8. Increment time by Δt: $t = t + \Delta t$

9. Go back to step 4 and repeat as often as desired.

Other steps may be used in more complicated versions of this process, for example to control the temperature of pressure of the system [11–15]. The potential energy between the particles is defined by the so-called force field. A typical modern force field has the form:

$$\begin{aligned} E_{tot} = E_{bonded} + E_{nonbonded} &= [E_{bonds} + E_{angles} + E_{torsions}] + [E_{VDW} + E_{Coulomb}] \\ &= [\sum_{bonds} k_b(r-r_0)^2 + \sum_{angles} k_a(\theta-\theta_0)^2 + \sum_{torsions} V_n(1+cos(n\phi-\delta))]] \\ &+[\sum_{i<j} \epsilon_{ij}[(\sigma_{ij}/r_{ij})^{12} - \sigma_{ij}/r_{ij})^6] + \sum_{i<j} q_i q_j/(4\pi\epsilon_0 r_{ij})] \end{aligned} \quad (1.16)$$

with the bonded terms being the harmonic potentials for the bond lengths r and the bond angles θ, and the Fourier series for the dihedral torsions ϕ. The non bonded terms correspond to the Lennard-jones potential describing the van der Waals force and the Coulomb potential describing electrostatic interactions through space. Parameters used in the force field are constantly adjusted based on quantum-mechanical calculations and empirical data from NMR spectroscopy [16–19].

The two non bonded terms account for the majority of the computational cost because all possible interactions of all atoms have to be taken into account, no matter how far away from each other they are. Cutoffs are usually used for these terms, that assume all forces between atoms over a particular threshold to be zero. Because this is somewhat problematic for the very long range Coulomb interaction, a method called particle mesh Ewald summation is usually used for electrostatic interactions [20, 21].

While MD simulations have become a very reliable tool over the years, it is still limited by two main factors:

1) Approximate force field parameters, which lead to inaccurate behavior of the simulated system,

2) Computational limitations and thus poor sampling of conformational landscape of the system.

These two limitations can lead to poor correlations when compared to NMR derived data. In practice, it is difficult to tell if poor correlations with experiments are due to experimental errors, poor force field parameters, sampling limitations or a misunderstanding of the values that are compared between the two methods, in particular because not all NMR experiments are sensitive to the same time scales. It is thus crucial to understand the origin of the discrepancies between MD derived and NMR derived data. This will be addressed in chapter 3. In chapter 5 on the other hand, we will use MD simulations to gain structural and mechanistic insight into role the N-terminal arm of the *Drosophila* Hox protein Scr plays in DNA binding specificity.

Chapter 2

Materials and methods

2.1 Protein expression and purification

2.1.1 Construct design

2.1.1.1 HM-Exd

We received plasmids for full-length Exd and Hth HM-domain (Exd1-376 and Hth1-247 from now on simply Exd and HM) on pET vectors from the laboratory of Dr. Richard Mann (Columbia).

When expressed by itself, Exd was largely insoluble. Cotransformation of BL21(DE3) *Escherichia coli* cells with both plasmids resulted in strong Exd expression but weak HM expression. Again most of the Exd was insoluble. After trying combinations of different vectors, the only viable solution was to coexpress both proteins on the same pET-Duet vector. Expression levels were about 1 to 1 and both proteins were now soluble (HM presumably binding to Exd and keeping it in solution).

Because this rather large construct would be hard to work with using NMR, attempts were made to identify shorter constructs yielding a minimal viable complex. This was done using the input of secondary structure prediction software (e.g. PsiPred), proton-deuterium exchange (HDX) mass spectrometry data from the laboratory of Dr. Gaetano Montelione (Rutgers), as well as Keri Siggers' (Columbia) partial proteolysis data [22] (see chapter 5). All these approaches agree mostly about large unordered areas at the termini of both proteins. Interestingly, even though all three sets of data suggest that residues 200-247 of HM must be largely disordered, we found that region to be necessary for solubility of the complex. After shortening termini to different degrees on both proteins, we found HM79-247 and Exd37-311 to be good candidates for subsequent studies (the combined construct will henceforth be named simply HM-Exd).

The tandem HM-Exd pET-Duet vector includes a His_6-tag N-terminal of the HM gene, followed by a TEV cleavage site.

2.1.1.2 Scr, AbdB and Exd homeodomains

Plasmids as well as glycerol stocks for the homeodomains of Scr and Exd were obtained from Nichole O'Connell (Columbia) [23]. Throughout this manuscript only two constructs were used for Scr: Scr 298-385 C362S and Scr 298-385 C362S N321D. Scr298-385 (C362S) comprises all of the homeodomain (HD) of Scr and includes a Cysteine to Serine mutation to prevent dimerization during the purification process. The double mutant Scr298-385 (C362S/N321D) has an additional Asparagine to Aspartate mutation meant to prevent the deamidation of Asparagine -3 (HD numbering), discovered by Nichole O'Connell during the work on her PhD thesis [23]. Both of these constructs were expressed from pET-15b plasmids generated by Dr. O'Connell using BL21(DE3)pLysS *Escherichia coli* cells.

The construct corresponding to the homeodomain of Exd discussed in chapter 5 (referred to as Exd320 in chapter 6) corresponds to Exd 238-320. This protein was expressed from a pET-15b vector using BL21(DE3)pLysS *Escherichia coli* cells in the same way as Scr. They both contain an N-terminal His_6-tag followed by a Thrombin cleavage site. The canonical homeodomain of Exd ends at residue 300, meaning that our constructs contain either 10 (Exd14) or 20 (Exd320) extra residues at the C-terminus. These extra residues are disordered in solution but become partly ordered upon binding of DNA *(vide infra)*.

The Exd construct described in chapter 6 (referred to as Exd14), as well as the construct for AbdB were provided to me by Dr. Nithya Baburajendran from the laboratories of Dr. Richard Mann and Dr. Barry Honig (Columbia) and Anna Kaczynska from the laboratory of Dr. Lawrence Shapiro (Columbia). Both genes are on a pDEST-HisMBP vectors that were created by Dr. Baburajendran using the Gateway^{TM} system. Both constructs contain an N-terminal His$_6$-tag followed by an MBP-tag (not used for our purposes) and a TEV cleavage site.

2.1.2 Protein expression

The expression protocol for unlabeled proteins was similar for all protein constructs described herein. For 1 liter of final expression culture medium, BL21(DE3) (for HM-Exd) or BL21(DE3) pLys (for all homeodomain constructs: Scr, AbdB and Exd) were inoculated in 3-5 ml Luria Broth (LB) in the presence of their respective antibiotics (50 μg/ml carbenecillin for all vectors described herein, plus 34 μg/ml chloramphenicol for constructs expressed in BL21(DE3)pLys cells, namely both Scr constructs, both Exd homeodomain constructs and AdbB). After becoming visibly cloudy (OD 0.6-1.0), cells were harvested by mild centrifugation (5 minutes at $<$ 3000g) and resuspended in 25 ml LB and left to grow overnight. In the morning cells were again harvested by mild centrifugation and resuspended in 1 liter of LB for final growth and expression. For all steps described here, the medium was supplemented with the appropriate antibiotics. In this final phase of cell growth, the optical density at 600 nm was measured at regular intervals ($\sim$20-30 min) until an OD of about 0.6 was reached, at which time expression was induced by adding 0.5

mM final concentration of IPTG (Isopropyl β-D-1-thiogalactopyranoside) to the medium. A final cell harvest was performed about 4 hours after induction (10 minutes at 5000 g) and cell pellets frozen at -20°C until they were needed for purification. All growth and expression was done at 37°C while shaking at 215-250 rpm.

2.1.3 Isotope Labeling

Isotope labeling requires a slightly changed expression protocol. The protocol for expression of backbone ^{15}N and perdeuterated protein was largely adapted from Gardner & Kay [24]. As opposed to unlabeled proteins, expression needs to happen in minimal medium supplemented with NMR active Carbon, Nitrogen or Hydrogen (Deuterium) source molecules. Since cell doubling times are longer in M9 medium (about 75 min versus about 20 minutes for rich medium) and even longer in D_2O based M9 medium (about 130 minutes), the total time of the expression protocol is increased.

In general, we made sure to stay within an OD range of 0.03 to 0.8 during the entire process. The protocol varied for different proteins but the general scheme was as follows: 4 ml of rich LB medium supplemented with the necessary antibiotics was grown over night at 25°C to OD 0.5-0.8 (temperature reduced as compared to unlabeled expression so that cell do not reach too high optical densities). In the morning cells were harvested by mild centrifugation as described above but then diluted in 50 ml M9 minimal medium made with H_2O. Cells were then grown during the day at 37°C until they again reached OD 0.5-0.8, when they were harvested by mild centrifugation and resuspended again to OD 0.03-0.1 in 1 liter M9 minimal medium. If a deuterated sample was desired this step was done in M9/D_2O instead of M9/H_2O. Because transfer to D_2O can be quite severe of a shock to the cells, often an additional step of 200 ml M9/D_2O growth was introduced between the 50 ml and the 1 liter steps for the cells to adapt to the D_2O environment before reaching the final expression volume. Protein expression was then induced with 0.5 mM final concentration of IPTG as before. If D_2O was used, expression was generally done over night at 18°C, whereas expression in H_2O was usually done for about 4-5 hours at 37°C. For ^{15}N backbone labeled protein M9 medium was supplemented with $^{15}NH_4Cl$, and unlabeled glucose. No ^{13}C backbone labeling was performed in this study, but can be achieved by using ^{13}C-labeled glucose instead [24] .

For the methyl labeled samples, the same protocol was used except that unlabeled NH_4Cl was used and ^{13}C-labeled ILV-precursors were added about one hour before induction. We used 100 mg of methyl labeled α-ketoisovalerate (for Valine, Leucine labeling) and 50 mg of α-ketobutyrate (for Isoleucine labeling). More information on this labeling scheme can be found in the literature [24–28].

2.1.4 Protein purification

Protein purification followed these general steps:

1) Lysis and sonication of cell pellet

2) Spinning down of cell debris

3) Affinity chromatography purification using HisTrap columns on ÄKTA protein purification system

4) Proteolytic tag cleavage and removal of tag and protease using HisTrap or HisTrap-Benzamidine (in tandem) columns

5) Ion exchange chromatography using HiTrap SP sulfopropyl cation exchanger columns

6) Optionally a step of gel filtration (size exclusion) chromatography before and/or after complex formation using Superdex S75 or S200 columns depending on the molecular weight in question.

For cell lysis, the frozen cell pellet was resuspended in 5ml per gram cell pellet of lysis buffer (see table 2.1). After resuspension we added one tablet of EDTA-free Roche complete mini protease inhibitor for up to 15 g pellet, 5mg per gram of pellet lysozyme and one spatula tip of DNase I powder. After about 30 minutes on ice with regular stirring, the lysate was sonicated for total time of 10 minutes in 20 second intervals with 50 second pauses in between (20 on / 50 off), until the lysate became clear. The lysate was then transferred to centrifugation tubes and insoluble cell debris separated by centrifugation at 30,000g for at least 45 minutes to one hour.

All purification steps were carried out at 4°C to minimize degradation and protease inhibitor was added when possible.

The supernatant was then carefully separated from the debris and applied to the Nickel charged HisTrap columns using a peristaltic pump. One 5 ml column was used for up to two liters of initial culture medium. After loading, the HisTrap columns were mounted on the ÄKTA purification system and washed with at least 10 column volumes (CV) of buffer A until the UV absorption was flat and back to the value seen for pure buffer A in the absence of the columns. A gradient was then programmed that went from 0 to 100% buffer B in about 50 minutes (2% per minute) to identify the start and end concentrations of buffer B needed to elute the protein from the column. In any subsequent purifications, this was replaced by a step wise elution as follows. The columns were washed at about 5-10% below the concentration of buffer B at which the protein starts eluting until the UV absorption becomes completely flat and then eluted with about 10% above the concentration at which the protein stopped eluting. Elute from this latter step was used for further purification. A subsequent wash of the column with at least 5 CV of 100% buffer B was performed, before rinsing with water and ethanol for storage of the columns at 4°C until the next time they were needed. Columns must then be rinsed again with water to remove the Ethanol before being equilibrated in buffer A. The elute containing the protein (as confirmed by SDS PAGE), was then concentrated using a centrifugal filter (usually Amicon® Ultra

Table 2.1: **Purification buffers**

Purification step (column type)	Buffer A (Equilibration)	Buffer B (Elution)
Affinity (HisTrap)	500 mM NaCl **20 mM Imidazole** 1 mM TCEP 50 mM TRIS pH 7.5	500 mM NaCl **500 mM Imidazole** 1 mM TCEP 50 mM TRIS pH 7.5
Ion exchange (HiTrap SP)	**100 mM NaCl** 1 mM TCEP 50 mM TRIS pH 7.5	**1000 mM NaCl** 1 mM TCEP 50 mM TRIS pH 7.5
Gel filtration (Superdex S75/S200)	100 mM NaCl (*) 1-5 mM TCEP (*) 10 mM HEPES or 50 mM TRIS (*) pH 7.5 (*) 0/50 mM $MgCl_2$ (*) (trace metal grade if used for NMR)	N/A (Equilibration and Elution buffers are identical)
Benzamidine elution buffer	N/A	500 mM NaCl 10 mM HCl pH 2.1
Crystallization buffer	200 mM NaCl 2 mM TCEP 10 mM TRIS pH 7.5 50 mM $MgCl_2$ (regular grade)	N/A

* Variable; depending on desired buffer conditions for the experiment. Ideally, no further buffer exchange would be needed and sample can simply be concentrated for the experiment.

10,000 NMWL for proteins over 10kDa; 3,000 NMWL for molecules under 10kDa, that is most of the DNA oligomers used as well as Exd14) and buffer exchanged into buffer A (concentrating to 1-2 ml, diluting into about 15 ml of desired buffer, followed by another concentration to 1-2 ml and one more dilution to about 15 ml, such that the original buffer has been diluted about 100 fold) of the cation exchange chromatography (”ion A”), for proteolytic cleavage and subsequent ion exchange chromatography. The appropriate protease was then added to the final diluted sample (10-15 ml). About 0.5 ml of TEV (1 mg/ml glycerol stock) protease or 20 μl of a 1U/μl bovine thrombin stock solution were added per up to two liters of original culture. A 15ml falcon tube containing the sample was placed in a 4 liter room temperature water bath for at least 6 hours. If the sample was left over night, the entire water bath was placed at 4°C so that the reaction would slow down over time. The amount of protease as well as the reaction time were adjusted according to the efficacy of the protease and the amount of protein that was expressed. For that end small aliquots are taken at regular intervals (30-60 minutes) for SDS PAGE to confirm efficacy of proteolysis so that subsequent purifications can be adjusted accordingly.

After proteolysis is completed and confirmed by SDS PAGE, the sample is again applied to the HisTrap column, which was reequilibrated in buffer A. This step is meant to remove uncleaved protein as well as cleaved His-tags and protease. The TEV protease used was His-tagged and will stick to the Nickel charged resin together with uncleaved protein and cleaved His-tags. Cleaved protein will not stick during this step. In the case of thrombin, which was not His-tagged, a benzamidine column was mounted in tandem with the HisTrap column. Benzamidine is a competitive inhibitor of Serine proteases and will bind Thrombin but not the protein of interest. Elution is performed exactly as described before except that this time the protein of interest will come off the colum at a much lower concentration of buffer B (at 0% for many proteins, but higher for some proteins who have some basic affinity to the Nickel resin). The fractions containing the protein of interest are again concentrated using a filter unit. The columns are again fully rinsed with >5 CV buffer B, water and ethanol for storage. The benzamidine column is eluted completedly using at least 5 CV of benzamidine elution buffer (see Table 2.1) before rinsing with water and ethanol for storage.

Concentrated protein was again buffer exchanged to ion exchange buffer A ("ion A") in the same way as described above for Nickel buffer A and applied to a cation exchange column (1x5ml HiTrap SP) with a peristaltic pump in the same way as done before in the case of the HisTrap columns. The ion exchange chromatography was done much in the same way as the affinity chromatography. An initial slow gradient elution was performed for each protein going from 0 to 100% buffer B in about 50 minutes to assess the elution profile of the protein. Subsequent elutions were done with a step wise gradient as described for the affinity chromatography. Fractions containing the protein were concentrated and buffer exchanged to the buffer of interest for experimentation. If an additional step of gel filtration was desired (for example to separate complexes from monomers, see below), no buffer exchange was performed but the sample concentrated to about 0.5 ml and applied to the desired gel filtration column, pre-equilibrated in the desired buffer (usually the buffer desired for the subsequent experiment, containing 50 mM $MgCl_2$ in the case of DNA-protein complexes).

When gel filtration was performed, we used the Superdex S75 column for all proteins and complexes of up to 30 kDa, and the S200 column otherwise. The column was equilibrated with the buffer of interest and the sample applied through the injection loop. In general a better separation is achieved with a slower flow rate and a larger column. We generally used a flow rate of 0.5-1.0 ml per minute (adjusting for back pressure according to the maximum allowed values for each column type as per manual).

2.1.5 Purification difficulties and protein stabilities

The homeodomains of Scr and AbdB express and purify very well but are extremely prone to proteolytic cleavage at the N-terminus, more specifically right at the beginning of the homeodomain. After a few days at 4°C most of the sample will have lost its NTA/linker region (according to Edman sequencing now starting at Arg3 of the HD), which of course will make studying that region with NMR difficult. HM is very prone to degradation as well, as mentioned before. N-terminal sequencing and Mass spectrometry of degradation products suggest that all of them start at Ala85 (only 6 residues are cleaved off), while still having different masses, which probably correspond to different cleavages at the C-terminus (predicted to be disordered, but which we had to include in the construct for solubility reasons - see above). Leaving HM-Exd at room temperature or 4°C for a few days will lead to precipitation, supporting the hypothesis that the C-terminus is needed for stability/solubility.

The HM-Exd complex cannot easily be frozen and thawed to keep it from degrading between purification and the NMR experiment, because this seems to lead to precipitation as well. Experiments thus had to be carried out directly after purification of this complex, to prevent proteolysis.

2.1.6 DNA preparation

DNA for NMR experiments was ordered as single strands from Keck Oligo (cartridge purified). DNA for X-ray crystallography was ordered as a preformed duplex from IDP (PAGE purified). Single stranded DNA was dissolved in water or the buffer of interest and buffer exchanged with a centrifugal filter of the appropriate molecular weight cutoff (generally 3,000 NMWL) against at least 12 ml of the desired buffer to remove impurities of molecular weight below the cutoff (shorter DNA strands). DNA concentration was then measured by UV absorption. Each strand was then mixed at equimolar amounts with its reverse complement strand, which was similarly prepared. The mix was then vortexed, spun down and heated to 95°C to 5 minutes. The tube was then left at room temperature for 10 minutes for the strands to anneal and form a duplex. A short step of centrifugation was used to eliminate condensation at the top of the tubes and concentration of the duplex measured again by UV absorption before mixing with proteins for complex formation.

2.1.7 Complex formation

All components of the complex were buffer exchanged to the desired final buffer for the experiment, containing 50 mM $MgCl_2$, needed to avoid precipitation, and kept on ice. For the NMR experiments this would usually be the gel filtration buffer described in table 2.1 including 50 mM $MgCl_2$, for X-ray cyrstallgraphy this would be the crystallization buffer described in the same table. DNA and Hox (Scr or AbdB) as well as Exd (Exd14, Exd320 or HM/Exd) were then prepared in ratios 1.2:1:1 for NMR and X-ray crystallography (or 1.2:1, if only one protein was present), unless we performed NMR experiments solely on the DNA in which case the ratios were closer to 1:1:1 (1:1, if only one protein present), so as to not contaminate our spectrum with signal from unbound DNA.

The first protein (generally Hox) was then pipetted in incremental volumes to the DNA to monitor for precipitation, while constantly being mixed by mild pipetting. Once the DNA and Hox samples were fully mixed, they would be left on ice for about 20 minutes before the second protein (usually Exd homeodomain or HM/Exd) was added in the same incremental fashion while monitoring for precipitation. Samples were kept on ice and bubbles from pipetting avoided at all times. After all components have been added the sample was again gently mixed and left on ice for at least an hour before being used directly for the NMR experiment or crystallization. Sometimes the complex was run over an additional size exclusion column to separate unbound protein or DNA and impurities, but because of the sensitivity to degradation, this step was usually skipped in favor of starting the experiment quickly after complex formation.

2.2 Nuclear magnetic resonance spectroscopy

2.2.1 Sample preparation

Almost all experiments on blue16, Scr, Exd, HM/Exd or complexes of these components, were done in a variant of the gel filtration buffer described in table 2.1. Most commonly this was: 100 mM NaCl, 1 mM TCEP, 50 mM TRIS, pH 7.5, 50 mM $MgCl_2$ (>99.995% purity, no paramagnetic trace metals), unless otherwise specified. Some experiments carried out on DNA alone were carried out in simple phosphate buffer at pH 6.0.

Each sample contained 10% D_2O for the purposes of locking the magnetic field strength and about 100 μM DSS (4,4-dimethyl-4-silapentane-1-sulfonic acid) for spectral referencing, which reduced the final concentrations of the solutes in the buffer by a factor of 0.9 (90 mM NaCl, 45 mM TRIS etc.). Additionally for most experiments with proteins we added protease inhibitor (usually Roche complete mini, EDTA-free if the sample contained magnesium).

2.2.2 NMR experiments

In the following a quick overview of the experiments that were collected in the context of chapter 5.

All experiments were collected on Bruker magnets, either in-house or at the New York Structural Biology Center (NYSBC). Spectra were processed either in Topspin (Bruker) or NMRPipe [29]. Table 2.2 lists the experiments and spectrometers used for the different

Table 2.2: **NMR experiments**

Experiment	Sample	Spectrometer
Jump return	blue16	500 MHz, 900 MHz
^{1}H NOESY	blue16	500 MHz, 900 MHz
Hahn Echo Jump return [30] (salt titration)	blue16	800 MHz
	blue16+ScrHD	800 MHz
	blue16+ExdHD	800 MHz
	blue16+ScrHD+ExdHD	800 MHz
	blue16+HM/Exd	800 MHz
	blue16+ScrHD+HM/Exd	800 MHz
$^{15}N/^{1}H$-HSQC/TROSY	$^{15}N/^{1}H$-ScrHD	600 MHz
	$^{15}N/^{1}H$-ScrHD+blue16	600 MHz
	$^{15}N/^{1}H$-ScrHD+blue16+ExdHD	600 MHz
	$^{15}N/^{1}H$-ScrHD+blue16+HM/Exd	600 MHz
$^{13}C/^{1}H$-methyl TROSY (HMQC)	$^{13}C/^{1}H$-ScrHD	600 MHz, 800 MHz
	$^{13}C/^{1}H$-ScrHD+blue16	600 MHz
	$^{13}C/^{1}H$-ScrHD+blue16+ExdHD	600 MHz

experiments.

2.3 X-Ray crystallography

2.3.1 Crystallographic screening

Initial screens for AdbB-Exd-DNA complexes were based on preliminary work by Nithya Baburajendran and Anna Kaczynska. DNA oligomers screened for this study were the ”magenta” 14mer GCATGATTTACGAC and the ”black” 14mer GCATGATAAATGAC.

Both 14mers were blunt ended. We made custom screens using the sitting drop method on 96-well plates (Axygen or Art Robbins Instruments). DNA and proteins were prepared as described above and equilibrated in the crystallization buffer described in table 2.1. The complex components were mixed to a final ratio of 400:400:480μM (AbdB:Exd14:DNA) and left for at least one hour on ice as descibed before. Screened precipitants were PEG 3350 and PEG 4000 (ranging from 15 to 26% w/v), a pH range from 5.3 to 9.8 and a $MgCl_2$ (Hampton Research) range of 0 to 300 mM. It should be pointed out that the initial buffer of the complex contained 50 mM $MgCl_2$, before being mixed with these screening buffers. Thus no drop was truly at 0 mM $MgCl_2$. 100 μl total of each crystallization condition was transferred to the wells of the 96-well plates and crystallization experiments were set up using the Mosquito® crystallization robot (TTP Labtech), which mixed 100 nl of sample with 100 nl of each crystallization condition into a sitting drop.

The trays were then left at 20°C in the Rock Imager (FORMULATRIX) and pictures taken according to a Fibonacci series of time intervals. Most crystals grew within 1 to 7 days and were harvested and mounted on 50 - 200 μm nylon loops mounted on metal bases (Hampton Research). 1 μl of cryo protectant (well solution plus 30% v/v glycerol) were pipetted onto each 200 nl drop prior to harvesting. The crystals were then flash frozen with the loops in liquid nitrogen and stored in vials (Hampton Research) under liquid nitrogen.

The conditions of the diffracting crystals are shown in table 2.3.

2.3.2 Data collection

Diffraction data was collected at the Advanced Photon Source (APS) at Argonne National Laboratory (Argonne, Illinois, USA) on beamline ID-24E. For both crystals, a total number of 200 images were collected with an angle increment of 1° and an exposure time of 1 second using a ADSC CCD Quantum 315 detector (Area Detector Systems Corporation). The

Table 2.3: **Crystallization conditions**

Oligomer name	Oligomer sequence	Final condition
magenta14	GCATGATTTACGAC	22% PEG 3350 90 mM $MgCl_2$ pH 9.0 (100 mM TRIS), 30% v/v glyercol (cryo)
black14	GCATGATAAATGAC	25% PEG 3350 0 mM $MgCl_2$(*) pH 5.3 (100 mM NaCitrate), 30% v/v glyercol (cryo)

* The well solution contained no $MgCl_2$, but the original protein sample contained 50 mM $MgCl_2$, making the actual concentration closer to 25 mM.

wavelength was 0.98 Å and the transmission 19.34% in both cases.

2.3.3 Data integration, model building and refinement

Collected images for both datasets (magenta14 and black14) were processed by RAPD (Rapid Automated Processing of X-ray Data, `https://github.com/RAPD/RAPD`) using the XDS software package [31] but the black14 dataset was reprocessed manually with iMosflm and merged, scaled and truncated to 2.4 Å using Scala from the CCP4 software package [32]. Phases were generated by molecular replacement using the program Phaser [33] from the CCP4 suite, using the structure of AbdB and Exd14 in complex with red14 (kindly provided by Dr. Baburajendran, unpublished) as template. Iterative cycles of building and refinement were conducted using Coot and Phenix.

Data collection and refinement statistics are summarized in table 6.2.

2.4 Molecular dynamics simulations

This section was published, in part:

Zeiske T, Stafford KA, Friesner RA, Palmer AG (2013) Starting-structure dependence of nanosecond timescale intersubstate transitions and reproducibility of MD-derived order parameters. Proteins: Structure, Function, and Bioinformatics 81: 499 - 509. [34] Reprinted with permission from John Wiley and Sons.

2.4.1 System preparation

This section describes the setup for the MD simulations in chapter 3 and chapter 5. The simulations described in chapter 4 have been carried out by Dr. Kate Stafford, and have been described in her thesis [35] and in the literature [36–39].

2.4.1.1 Simulations of GB3

Set A starting structures were derived from the 1.1Å X-ray crystal structure (PDB code 1IGD) [40] with the N-terminus altered as described previously [41] to recapitulate the construct used for spin relaxation studies [42,43]. (D1-5, T6M, T7Q; mutations performed in PyMOL [44]). Side-chains conformations were further optimized using PLOP [45]. Starting structures for the set B simulations were derived from the X-ray crystal structure (1IGD) in the same fashion, but independently from set A and without PLOP optimization.

All structures, except reruns of completely solvated systems of previously conducted simulations [41] (set A), were prepared for simulation in Maestro [46] with the Maestro Protein Preparation Wizard, which also added hydrogen atoms to the structures. Maestro was used to solvate the protein with TIP3P (or TIP4P if specified) water molecules [47] in cubic boxes of 50, 54, 75, or 90Å edges (all lengths 1Å). This corresponds to minimum buffer layer thicknesses of 1 nm (50 or 54Å depending on orientation of the molecule in the box), 2 nm (75Å), or 3 nm (90Å). The system was either neutralized with two sodium ions or an additional 0.15M NaCl was added as specified.

A summary of differences between the original set A and set B simulations can be found in Table 3.1. Reruns of completely solvated set A simulations were processed with Ptraj from the AMBER9 suite [48] and then prepared in Maestro.

2.4.1.2 Simulations of Scr and its mutants with fkh and fkhCON

Simulations of Scr and its mutants with fkh and fkhCON were prepared in much the same way as simulations of GB3.

All starting structures were derived from the two crystal structures 2R5Y and 2R5Z described in [49]. The Scr protein moiety for all simulations was derived from the crystal structure with the specific *in vivo* target fkh (2R5Z), because it included Arginine 3. Its linker region, as well as Exd and any ions and water molecules were completely removed, such that the construct started at Arginine 3 for its N-terminal residue. While introducing a charged NH_3^+ group at this position could certainly cause the simulation to behave in a non-native way, we preferred it over trying to model further residues into the structure whose positions we knew nothing about. Since we wanted to learn about qualitative differences for the behavior of the N-terminus between proteins with point mutations at positions 4 and 6, and had to choose between two different kinds of biases, we decided introducing this charge and leaving out non resolved residues would suffice for our purposes.

For simulations with fkh, we included the coordinates of fkh from the same PDB structure 2R5Z. For simulations with fkhCON we modeled Scr from 2R5Z onto Scr from 2R5Y by RMSD minimization and saved its coordinates together with the coordinates of fkhCON from 2R5Y. Missing atoms were added, phosphates added to the DNA as needed and any manual mutations were then carried out in Maestro [46], where the systems were also solvated and prepared for simulation as described above. Proteins were solvated in a cubic box of 1nm minimal buffer layer using TIP3P water molecules [47] and 0.15 mM NaCl were added to the system.

2.4.2 Simulations

AMBER ff99SB (or ff99SB-ILDN as specified) simulations were conducted with the Desmond MD software package (Academic Release 3) [50]. Particle-Mesh-Ewald periodic boundary conditions were used with a 9 or 12Å cutoff for electrostatic interactions as specified. The integrator used time steps of 2 fs. Bonds to hydrogen atoms were constrained using the M-SHAKE algorithm [51]. Energy minimization (convergence threshold 1 kcal/mol/Å) and a 1 ns NPT equilibration simulation at 297 K and 1 atm were conducted for each system. Starting structures for production runs were extracted at equally spaced time intervals from the second half of the equilibration runs. Production runs were conducted for 2.4 ns at constant volume and energy conditions (NVE). Coordinates were written out every 1 ps.

2.4.3 Trajectory analysis

All trajectories were analyzed using VMD [52]. The effects of overall tumbling during the simulation were removed by superposing the C_α atoms of each frame by RMSD fit to the first frame of the simulation. Orientational autocorrelation functions for the NH bond vectors were calculated as described previously [41]. If no convergence threshold was applied, then order parameters were calculated as described previously [41,53]. To judge convergence, the last value of the autocorrelation function [C(τ=1200 ps)] for a given residue and simulation was compared to its mean value in the middle region (frames 300−900 ps). If the absolute difference was within the threshold (set to 0.005), then the order parameter was set to the mean value of the autocorrelation function in the middle region (300−900 ps) and included into averaging over simulation blocks. If the difference was larger than the defined threshold, then these data were excluded from order parameter averages. Simulated order parameters were scaled by $\xi = (1.02/1.04)^6 \approx 0.89$ for comparisons with spin relaxation derived data to account for zero point vibrational motions of the NH bond vectors [54].

Chapter 3

Reproducibility of molecular dynamics derived order parameters

This chapter was published, in part:

Zeiske T, Stafford KA, Friesner RA, Palmer AG (2013) Starting-structure dependence of nanosecond timescale intersubstate transitions and reproducibility of MD-derived order parameters. Proteins: Structure, Function, and Bioinformatics 81: 499 - 509. [34] Reprinted with permission from John Wiley and Sons.

3.1 Introduction

Their atomistic detail and high resolution in both space and time make molecular dynamics (MD) simulations an ideal tool for studies of the conformational dynamics of biological molecules, especially in synergy with experimental methods such as NMR. The insights obtained from such joint investigations are necessarily limited by deficiencies in simulation procedures that reduce quantitative agreement with experimental data. Despite efforts made over the last decade, discrepancies persist between different MD simulations of the same system and between MD simulations and NMR [41].

The focus of the project described in this chapter is to identify sources of inaccuracies in MD simulations, both in comparing multiple simulations to each other and in comparing simulations to NMR spin relaxation data. We use the B3 immunoglobulin-binding domain of streptococcal protein G (GB3), a 56 amino acid α/β protein, which has served as a common model system for protein dynamics. As in many other investigations [41, 55, 56], the square of the generalized order parameter, S^2 (henceforth simply called the order parameter) is used to describe the orientational conformational distribution of the backbone amide (NH) bond vectors. Order parameters can be derived experimentally from NMR spin relaxation rate constants or residual dipolar couplings (RDCs) [8]. A limitation of spin relaxation experiments is their insensitivity to conformational dynamics on timescales similar to or longer than overall rotational tumbling of molecules. In contrast, RDC-derived order parameters are sensitive to motions on a wider range of timescales up to milliseconds, but contributions from processes much slower than MD trajectory lengths are not captured by simulations [55, 57].

We compare three sets of simulations: set A is based on 14 trajectories previously reported [41], with additional trajectories that were produced herein from the same starting structures; set B consists of 16 trajectories recorded using starting structures generated independently from set A; and set C consists of a 1.2 microsecond simulation previously described in the literature [17]. Comparisons between these trajectories demonstrate that multiple simulations using a single force field (AMBER ff99SB) can result in discrepancies between the simulated order parameters because the choice of starting structures influences subsequent sampling. Even sets of starting structures that are indistinguishable by backbone RMSD measures can yield notably different dynamical behavior for GB3. For example, such differences arise from a single tyrosine hydroxyl proton with two orientations, owing to different methods of protonation that are unable to interconvert even during very long simulations. Other parameters of the simulation protocol, including box size or geometry, water model, salt content, or force cutoffs, have at most minor influences on the behavior of the system during simulations.

Sampling-related problems are reduced in longer simulations, but even microsecond simulations are still strongly dependent on the starting point on the conformational energy landscape. Many sampling-related discrepancies between simulations are consequences of nanosecond timescale motions, often related to sidechain rearrangements or breaking of hydrogen bonds (sometimes linked to water invasion), that lead to unconverged NH autocorrelation functions on the timescale of the analyzed simulation blocks, typically chosen to be of order of the rotational correlation time of GB3 for comparisons with NMR spin relaxation data. Applying a threshold to exclude simulation blocks whose autocorrelation functions fail to converge eliminates nearly all differences between simulations with different starting structures and yields order parameters that are in much better agreement with experimentally derived values.

3.2 Influence of Simulation Parameters

The initial set B trajectories and the previously published set A trajectories were generated as described in Chapter 2 (Materials and methods, summarized in Table 3.1). The calculated order parameters for both sets were compared with experimental order parameters derived from NMR spin relaxation measurements (Figure 3.1) [43]. As described previously [41], the MD-derived order parameters are underestimated compared to experimentally obtained order parameters, especially in the flexible loop regions and at the termini of the protein. In addition, the set B order parameters had major discrepancies with the set A simulations, mainly within the first two loops of the protein (Figure 3.1). In loop 1, the main differences are for residues Gly9 and Gly14. These two residues also show

Table 3.1: **Summary of the differences of the original set A and set B Simulations**

	Set A	Set B
Initial coordinates	1IGD	1IGD
Mutations	1IGD	1IGD
Protein preparation (e.g. addition of hydrogens)	Δ1-5, T6M, T7Q	Δ1-5, T6M, T7Q
PLOP optimization [45]	Yes	No
Water model	TIP4P [47]	TIP3P [47]
Force field	AMBER ff99SB [58]	AMBER ff99SB [58]
Salt	2 sodium ions	0.15 M NaCl
Solvent box	Orthorhombic, 1 nm minimal buffer layer	Cubic, 1 nm minimal buffer layer
Temperature	297 K	297 K
Length of simulations	2.4 ns	2.4 ns
Original number of trajectories	14	16

Note that the N-terminal alterations were performed independently for set A and set B.

large variances in the order parameters calculated for individual simulations within each set. In loop 2, the main differences are for residues Ala20 and Asp22. These two residues have rather large order parameters in set B that are closer to the experimental data.

As described below, a series of additional simulations were performed to identify sources of differences between the set A and set B simulations and between the MD-derived order parameters and experimental values.

Water model, box symmetry, and salt concentration were successively changed to eliminate differences between the simulation protocols used for set A and set B. Fourteen simulations of the set A starting structures using the TIP3P water model yielded order parameters that were indistinguishable from the original values obtained using TIP4P water

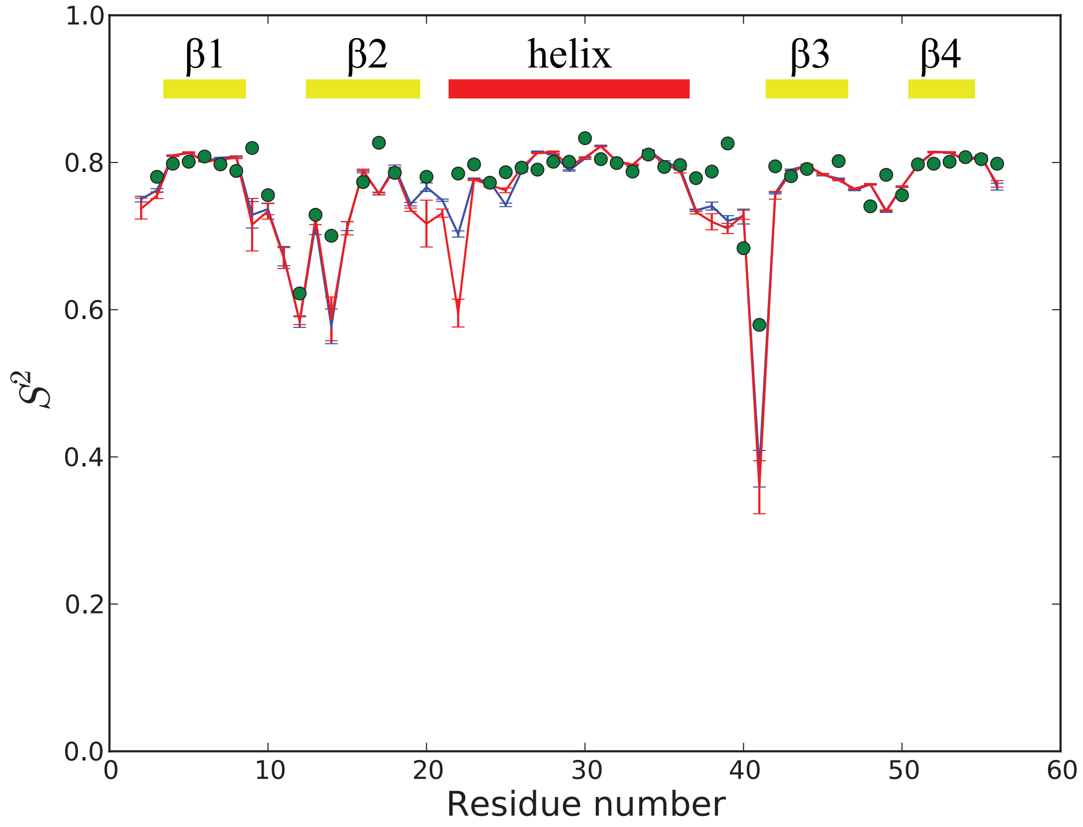


Figure 3.1: **Order parameters of both sets compared to NMR derived order parameters**
Order parameters for simulations using the set A (red) and set B (blue) starting structures. For comparison, the experimental values are shown as filled green circles [43]. The biggest discrepancies lie within loop 2 (Ala20 and Asp22), and to some extent at the N-terminus and loops 1 (Gly9) and 3 (residues 37 - 41). Error bars represent standard errors.

and the discrepancies with set B simulations were unaltered (results not shown).

Nine simulations of set B structures were performed in a 50Å cubic box (1 nm minimal buffer layer, with the long axis of the protein along the diagonal of the cube) using only two sodium ions to neutralize net charge, to correspond to the original set A protocol (Figure 3.2). In addition to expected minor differences in loop 1, the flexibility and variability of the N-terminus and Ala20 are slightly increased, similarly to the previously published results [41].

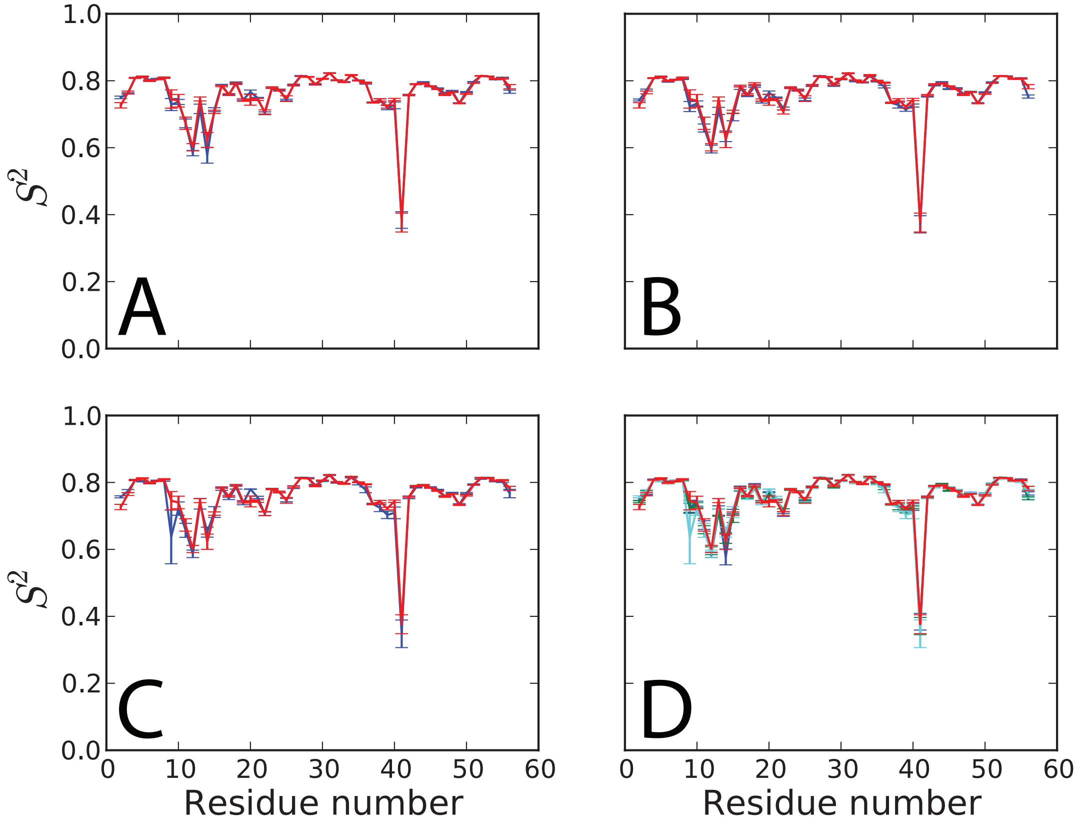


Figure 3.2: **Simulations with 0.15 M salt compared to simulations with only two neutralizing sodium ions**
The blue line in A), B) and C) shows the order parameters for simulations with 0.15 M salt for a 1 nm, 2 nm and 3 nm minimal buffer layer, respectively. The red line in A), B) and C) shows the order parameters for simulations with only two neutralizing sodium ions and a minimal buffer layer of 1 nm. D) Order parameters for the simulations with 0.15 M salt and all three box sizes (blue 1nm, green 2nm, cyan 3nm) compared to results for simulations with two neutralizing sodium ions and 1 nm buffer layer (red). Error bars represent standard errors.

Resolvating five of the set A starting structures according to the solvation protocol used for set B (TIP3P, 1 nm minimal buffer layer, 0.15M NaCl, cubic box) produced order parameters in agreement with the original set A trajectories (Figure 3.3), with a few nonconverged low order parameters for Ala20, suggesting a larger transition in those trajectories, and a number of lower order parameters for Asp22. The trajectories having a low order parameter for Ala20 also have a low order parameter at the N-terminus; the correlation between these two sites will be discussed below. The dynamics of Asp22 appears uncorrelated to the movements of Ala20 or the N-terminus.

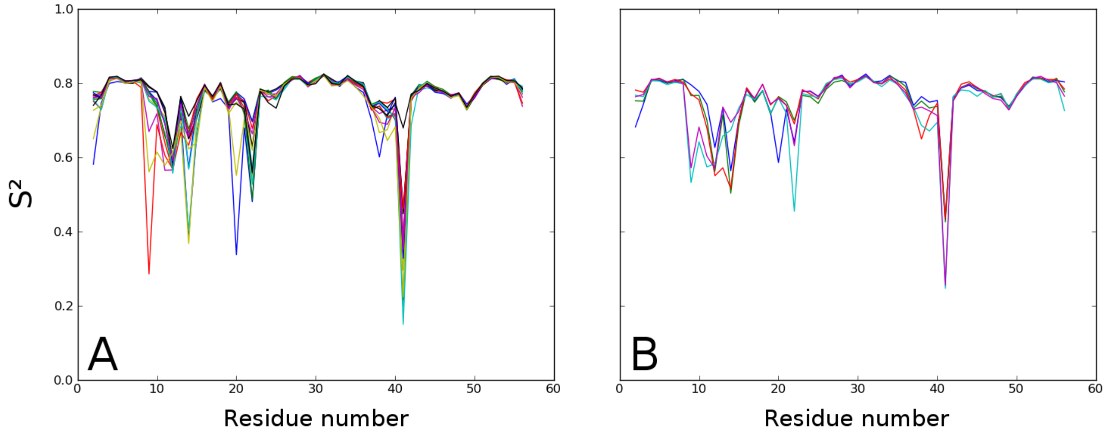


Figure 3.3: **Solvation protocol has only a minor influence on resulting order parameters**
A) 14 simulations resulting from rerunning starting structures derived from the 14 set A simulations in the original orthorhombic boxes including solvent and two neutralizing sodium ions.B) 5 simulations using 5 of these 14 starting structures after deleting all solvent molecules and ions and resolvating in a cubic water box with 0.15 M NaCl. Both A) and B) show increased flexibility in loop 2 (Ala20, Asp22) as in the original set A trajectories.

To study the influence of the water box size, the set B simulations were expanded to include 17 trajectories for a 75Å cubic box and five trajectories for a 90 Å cubic box, corresponding to minimal buffer layers of 2 and 3 nm, respectively. Order parameters were calculated for all trajectories and compared with the results for the 54 Å cubic box (Figure 3.4).

Agreement between order parameters obtained from the three different box sizes is very good. Only two outliers differences in order parameters > 0.05 were observed: Glycine 9

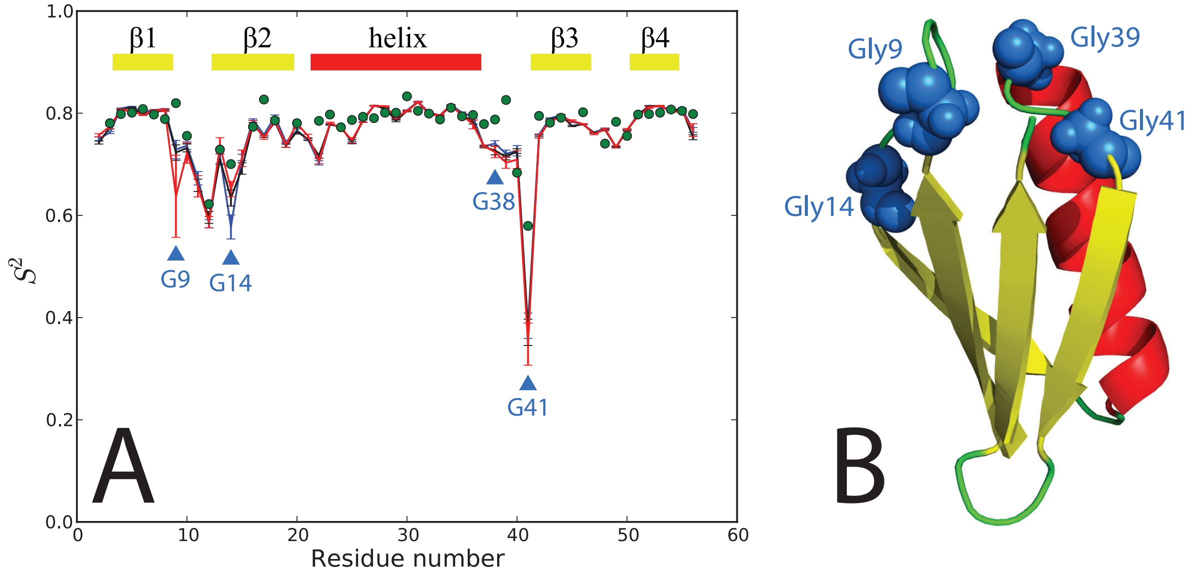


Figure 3.4: **Influence of the water box size is minimal**
Order parameters for minimal buffer layers of 1 nm (blue), 2 nm (black) and 3 nm (red) in cubic water boxes. Experimental values from spin relaxation measurements are depicted as green circles. Error bars represent standard errors. Glycines are marked with black triangles. B) Structure of GB3 and the positions of the four glycines.

(difference in order parameters of about 0.09 between 1 and 3 nm buffer layer sets) and Glycine 14 (difference in order parameters of about 0.07 between 1 and 3 nm buffer layer sets). Two more minor outliers were observed: Glycine 38 and Glycine 41 (to a certain extent affecting the intervening residues as well). All those residues are also outliers in comparison with the experimental data. Note that these are all four of the glycines found in GB3 and that they all lie at loop hinges at the ends of secondary structures (Figure 3.4). The peptide bond of Gly14 with Lys13, however, has such variability in the different structures that the β-sheet sometimes extends to Leu12. In the X-ray crystal structure, the sheet also extends to residue 12, but has a clearly visible kink at Gly14. These results suggest that glycine parametrization remains a weakness of current MD force fields [58,59]. Because all glycines in GB3 flank secondary structure elements, the problem may arise specifically for glycines in these special positions.

To test whether simulations of larger box sizes were affected by the choice of the electrostatic cutoff, we reran one of the 2 nm simulations of set B increasing the cutoff from 9 to 12. The results were not significantly different (results not shown).

The parameters for the sidechains of isoleucine, leucine, aspartate, and asparagine in the AMBER ff99SB forcefield have been reoptimized [17], yielding a new forcefield termed AMBER ff99SB-ILDN. We conducted fourteen 2.4 ns simulations using the set A starting structures and twenty-six 2.4 ns simulations using the set B starting structures together with the new forcefield. All simulations were performed using a 50Å cubic box with two sodium ions to neutralize the system (Figure 3.5). Simulations of the set A starting structures with the ff99SB-ILDN force field yield somewhat larger order parameters for Ala20 and the N terminus, but otherwise the discrepancies with the results from set B remain. Interestingly, flexibilities of loop 1 (residues 9 through 13) and to a minor extent loop 3 (residues 38 through 40) are more pronounced with the new force field for simulations using the set B starting structures. However, the increased flexibility in loop 1 may reflect sampling limitations, because reduced order parameters are not observed using a 1.2-μs trajectory, which was also conducted with the ff99SB-ILDN force field (*vide infra*).

3.3 Transitions in conformational space and sampling

The above investigations suggest that the starting structures are the main cause for the discrepancies in order parameters between the different sets of simulations. Backbone RMSD comparisons of starting structures from sets A and B to each other and to the X-ray crystal structure were less than 1.2Å in all cases, with no apparent correlation between RMSD and discrepant order parameters (results not shown). One of the main features of the dynamic behavior of Gly9, Gly14, and Ala20 is their apparent bifurcated behavior in all the simu-

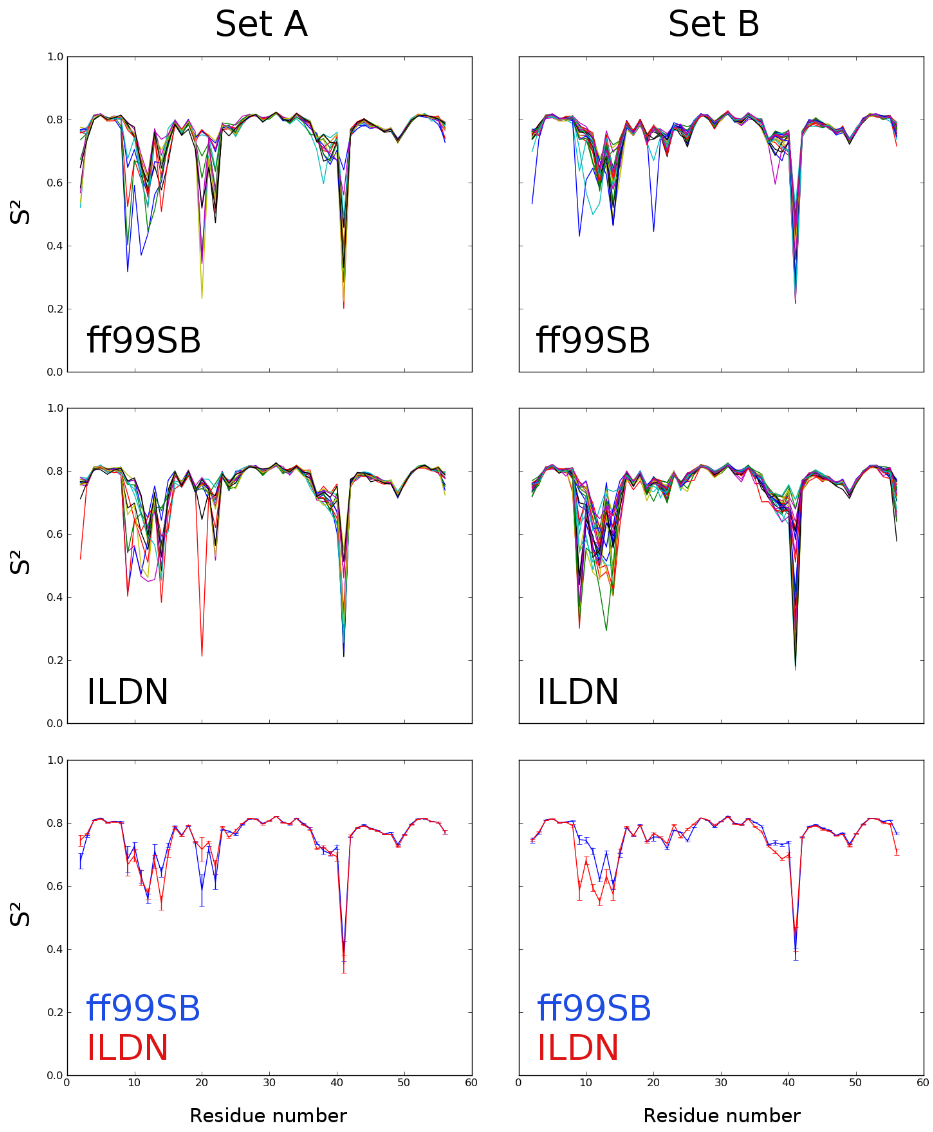


Figure 3.5: **Minor changes are seen after using the improved AMBER ff99SB-ILDN force field**

The left panels show simulations related to the set A starting structures, the right panel to the set B starting structures. The first row shows order parameters from simulations in the AMBER ff99SB force field, the second row shows order parameters from simulations in the sidechain-corrected AMBER ff99SB-ILDN force field, and the last row shows order parameters averaged over simulations for each set of starting structures and force field with error bars representing standard errors.

lations. The order parameter is either high (corresponding to a converged autocorrelation function) or low (corresponding to an unconverged autocorrelation function) but seldom adopts intermediate values (compare Figures 3.4 and 3.6). This behavior indicates that GB3 undergoes conformational changes on nanosecond or longer timescales and simulated results are therefore strongly dependent on sampling in the nanosecond-length trajectories. Asp22 shows a wide spread of order parameters in the set A simulations, but is rigid in all simulations in set B (Figures 3.4 and 3.6). Gly9 and Gly14 are seen to undergo transitions in both sets of simulations. Ala20 exhibits large-scale transitions predominantly in the set A simulations, although these occur occasionally in the set B simulations. In general, the set A starting structures populate a region of the energy landscape for which these conformational transitions are more probable. Backbone RMSDs of each frame from the first frame of each trajectory show transition-like behavior for trajectories that exhibit low order parameters for Gly9 and Gly14 (results not shown). The same transition-like behavior is observed for a similar backbone RMSD analysis of loop 1 alone, but only to a minor extent for loop 2 (results not shown). Thus, the conformational dynamics for Asp22 are more variable than for Gly9, Gly14 or Ala20. In addition, the occasional transitions of Ala20 are not concerted with transitions of Asp22 and do not affect a large portion of loop 2, in contrast to the effects on loop 1, where a number of interactions are rearranged in concert. In other words, loop 1 seems to make transitions as a unit whereas loop 2 does not (*vide infra*).

3.4 Increased sampling with a 1.2-μs long trajectory

We analyzed a 1.2-μs trajectory of GB3 (kindly provided by D.E. Shaw Research [17]) to assess the effects of greatly increasing the simulation time. Order parameters were calculated for 500 blocks of length 2.4 ns (set C), to allow comparison with the order

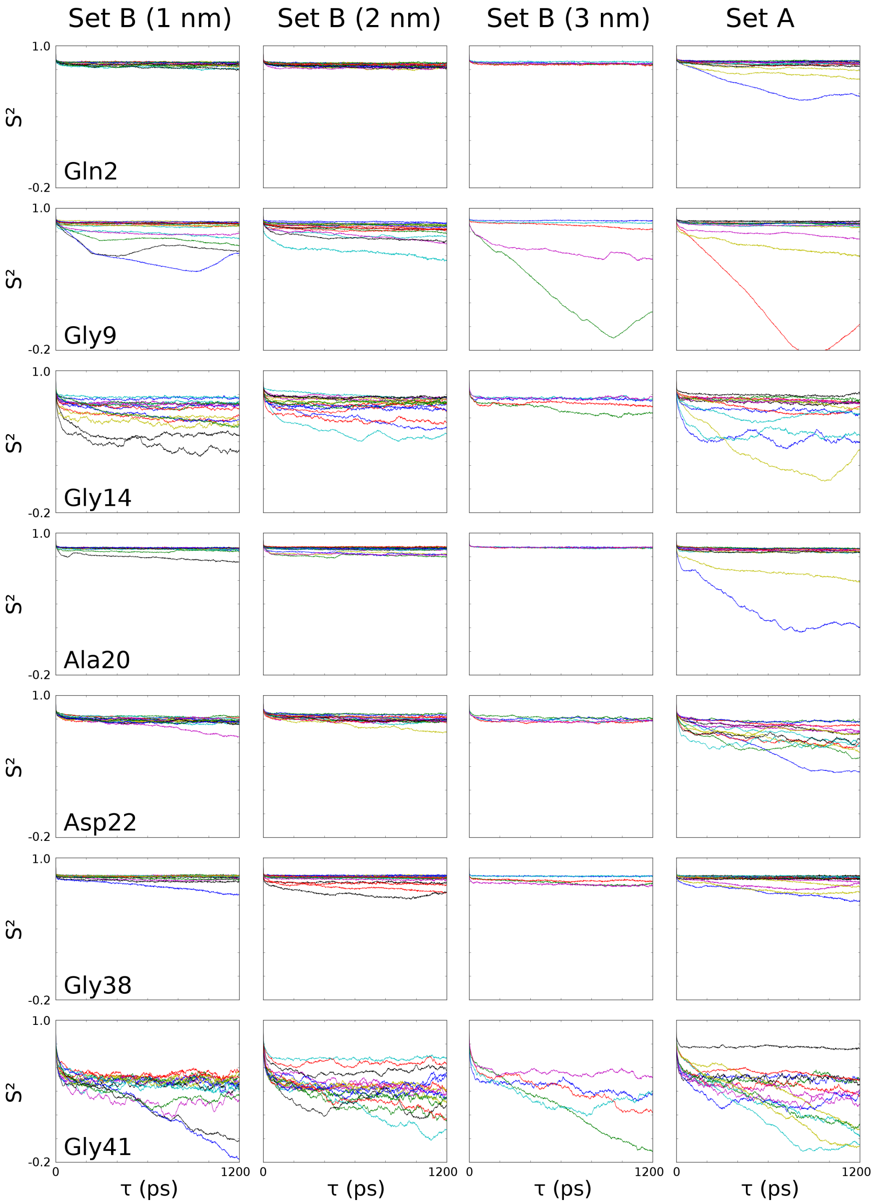


Figure 3.6: **Autocorrelation functions for Gln2, Gly9, Gly14, Ala20, Asp22, Gly38, and Gly41**

The first three rows show the simulations of the set B starting structures in cubic boxes with 1 nm, 2 nm and 3nm minimal buffer layers respectively. The fourth row shows the rerun of the starting structures derived from set A simulations. The bifurcated behaviour of converged versus non-converged autocorrelation functions is more pronounced for the set A starting structures especially for residues 2, 20 and 22. Also note that the aberrant trajectories for Gln2 and Ala20 in set A are identical (yellow and blue lines).

parameters obtained from the sets A and B simulations. The results are shown in Figure 3.7. A number of residues have large variances in order parameters, namely both termini, loop 1 (residues 9 – 15), Ala20, Asp40, and Gly41 and to some extent Gly38, Val39, and Phe30. With the exception of Phe30, those residues are all in loops, termini, or flank secondary structure elements.

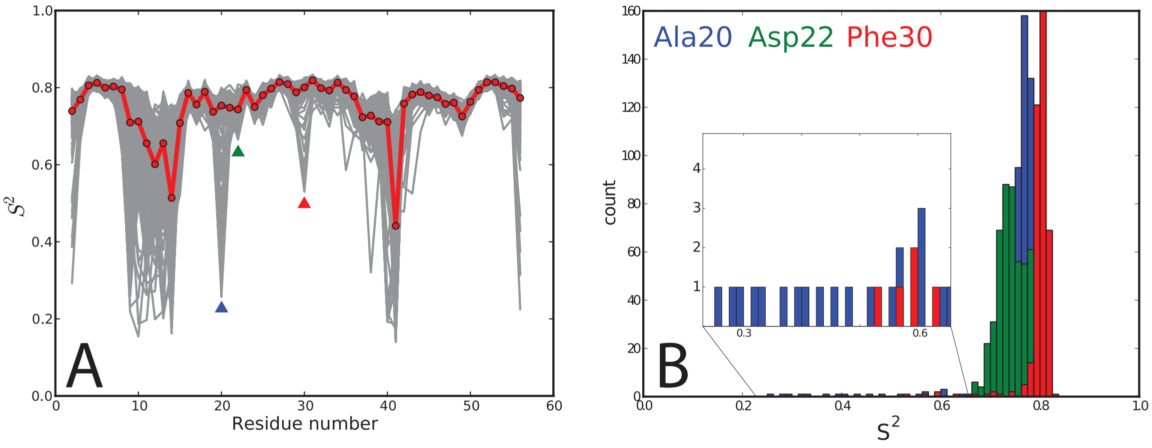


Figure 3.7: **Order parameters of the 1.2-μs trajectory**
A: Order parameters for all 500 2.4-ns blocks of the 1.2-μs trajectory (grey lines). The red line and dots indicate the average values, the error bars were omitted for clarity. Ala20, Asp22, and Phe30 are marked with a blue, green, and red triangle, respectively. Ala20 and Phe30 clearly show bifurcated behavior: the majority of the order parameters are high (corresponding to converged autocorrelation functions) with a number of very low order parameters, which do not affect the average over the large number of blocks. B: Distribution of order parameters for Ala20, Asp22, and Phe30. Most order parameters are high, but Ala20 and Phe30 exhibit a small number of excursions to very low order parameters (corresponding to unconverged autocorrrelation functions). Asp22 on the other hand does not show any outliers.

The results from the 1.2-μs trajectory show that the mean value of the order parameter for Ala20 is almost unaffected by infrequent fluctuations to conformational states with high local flexibility. Additionally, the low order parameters for Ala20 do not coincide with low order parameters for any of the other residues in loop 2. Thus, transitions to other subensembles appear to be sampled too frequently using the starting structures for set A. Phe30 also exhibits infrequent transitions to other conformational states, linked to low order parameters for those transitions. Examination of the trajectory shows that the transitions of Phe30, situated in the middle of the a-helix, is accompanied by a bending of the helix and other movements all along the length of the protein, most notably in the loops and termini (especially loop 1 and the C-terminus). This transition was not sampled in any of the 2.4-ns trajectories (sets A or B). Very strikingly, Asp22 remains constrained for the entire 1.2-ls trajectory.

3.5 Long range effects and the influence of sidechain conformations on local transitions

Conformational transitions of Ala20 are very strongly coupled to movements of the N-terminus, as shown by the correlation between order parameters for Gln2 and Ala20 over the 500 blocks derived from the 1.2-μs simulation (Figure 3.8). Two hydrogen bonds between the backbones of Ala20 and Met1 transiently break, allowing flipping out of the N-terminus and rearrangement of Ala20. A similar correlation is observed between residues in loop 1 and the C-terminus, as a result of fluctuating interactions between backbone and sidechains of residues 55 and 56 with the backbone and sidechains of residues 8 through 11. However, loop 1 also interacts with loop 3, leading to a lower correlation between the motions of loop 1 and the C-terminus.

Although the overall RMSDs from the crystal structure of the backbone of all the structures in sets A and B are similar, plots of the absolute differences for each C_α along the chain of the GB3 show that the starting structures from set A have a higher variability than structures from set B relative to the X-ray crystal structure (Figure 3.9A). Backbone

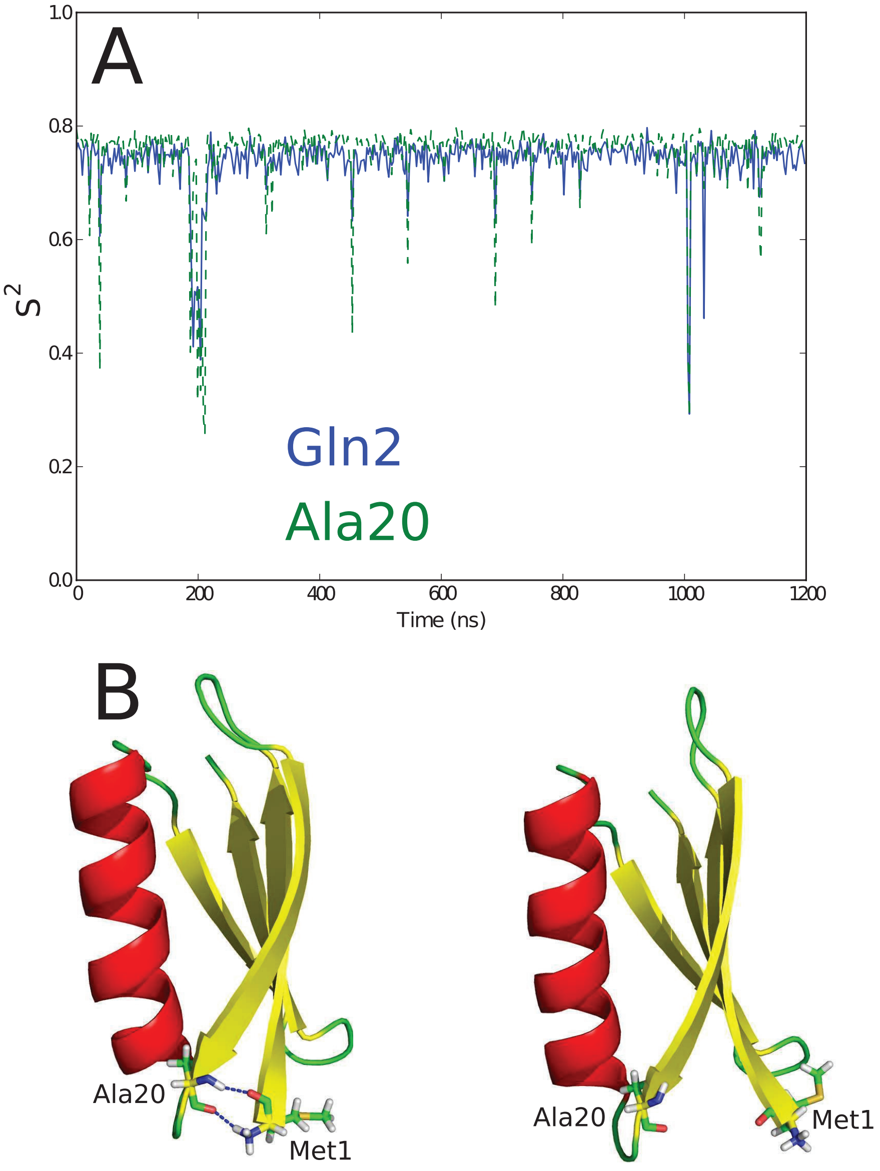


Figure 3.8: **The motions of Glutamine 2 and Alanine 20 are strongly correlated** A: Order parameters for Gln2 (blue solid) and Ala20 (green dashed) are strongly correlated over the 1.2-μs trajectory. B: Ala20 flips when its hydrogen bonds with the N-terminus are broken. The left panel shows the native hydrogen-bonded state. The right panel shows the state with the flipped-out N-terminus and broken hydrogen bonds.

RMSDs for set A and set B starting structures from the starting structure of the 1.2-μs simulation show a substantial deviation for set A at the C_α of Val21 in loop 2 as well as at the N-terminus (Figure 3.9B). The NH of Asp22 is part of the peptide bond to Val21, suggesting a source of different simulated properties for sets A and B. The sidechain of Val21 has a different conformation in the set A starting structures than in the set B structures or the starting structure of the 1.2-μs trajectory. Additionally, the backbone around Val21 is slightly displaced in set A starting structures, which explains the high C_α deviation of Val21 (Figure 3.9C). In the original X-ray crystal structure, the sidechain of Val21 is in the same conformation as in the set B starting structures, but the backbone around Val21 adopts an intermediate position between the two clusters of structures, explaining why the RMSDs of the Val21 C_α atoms from the PDB structure are similar for both sets A and B.

3.6 Differences in protonation influence backbone dynamics

As described previously [41], the carbonyl of Val21 forms a non-native hydrogen bond with the hydroxyl group of Tyr3 in set A simulations. Indeed, in the set A starting structures, this hydroxyl group is pointing towards loop 2 and Val21 (Figure 3.10A), poised to form the hydrogen bond. In set B and in the 1.2-μs trajectory, the hydroxyl group of Tyr3 is pointing away from loop 2. In this conformation the hyrodxyl group is hydrogen bonded to a water molecule, and is not available to hydrogen bond with Val21. Although Val21 undergoes sidechain rearrangements along the 1.2-μs trajectory, the hydroxyl group of Tyr3 never rotates to point towards loop 2 and thus never forms a hydrogen bond with Val 21 (Figure 3.10B). The different orientations of the hydroxyl group arise when preparing the system for simulation by adding hydrogen atoms to the crystal structure. Simulations

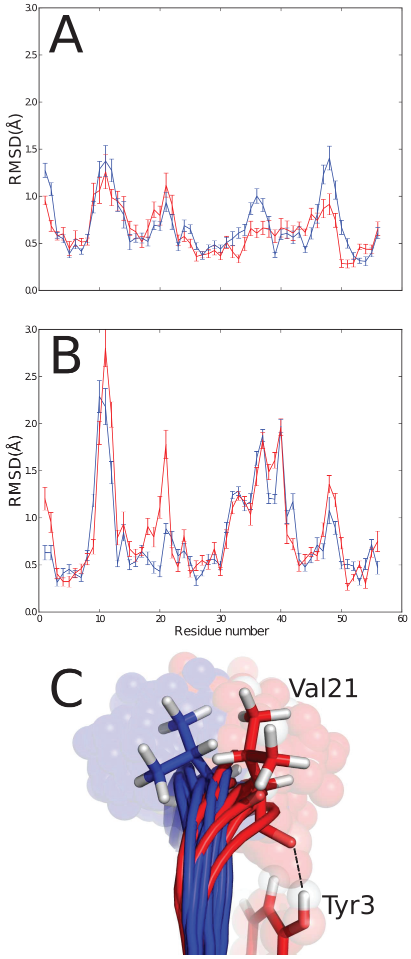


Figure 3.9: **Starting structure differences between sets A and B**
A: C_α RMSDs from the PDB structure averaged over all set A (red) and set B (blue) starting structures. No obvious differences in loop 2 are apparent. Error bars represent standard errors. B: C_α RMSDs from the starting structure of the 1.2-μs trajectory averaged over all set A (red) and set B (blue) starting structures. Clear differences are observed in the N-terminus and loop 2, most prominently at the C_α of Val21. Error bars represent standard errors. C: Position of the sidechain of Val21 in all set A (red spheres) and set B (blue spheres) starting structures. Ribbons represent the backbone of loop 2; sticks are two representative sidechain conformations of Val21 for each set of structures. In the case of set A, the backbone carbonyl of Val21 is pulled toward the hydroxyl group of Tyr3 (red sticks and spheres).

of the closely related protein GB1 also show increased flexibility of Asp22 and the Tyr3 hydroxyl group is oriented towards Val21 (unpublished results).

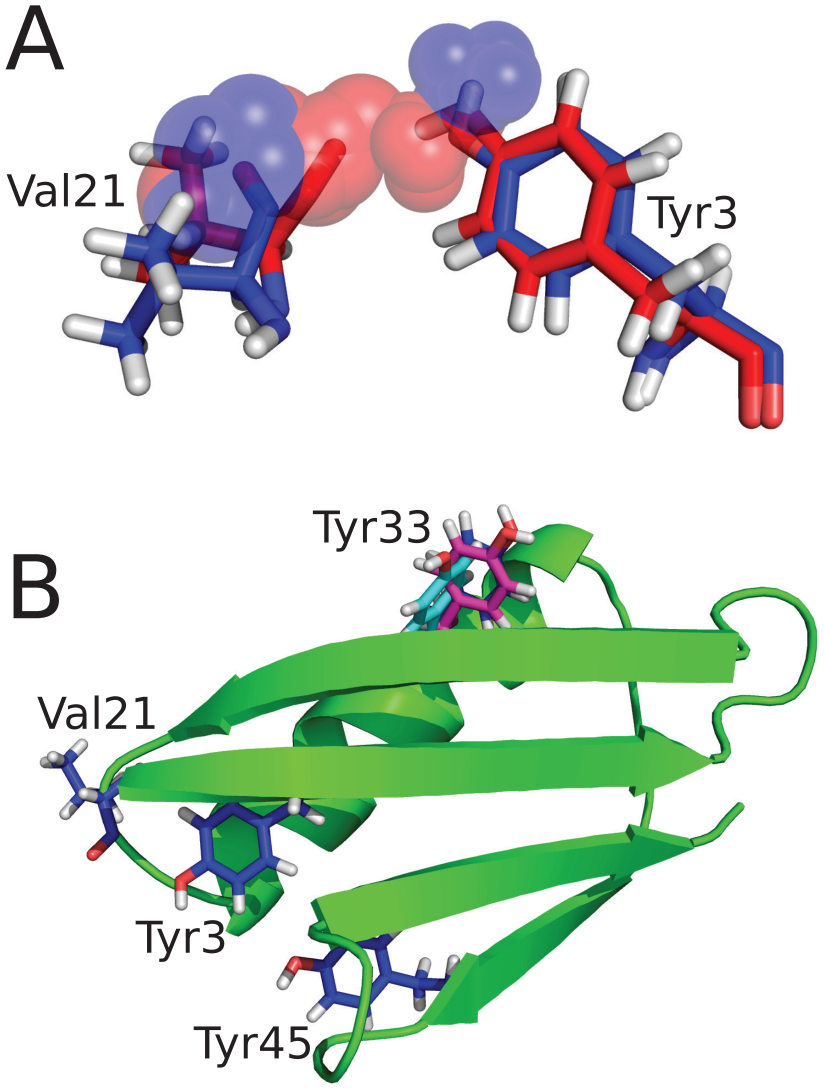


Figure 3.10: **Different orientations of the Tyrosine 3 hydroxyl group**
A: The hydroxyl group of Tyr3 points towards the backbone carbonyl of Val21 in all of the set A (red) but away from it in all of the set B (blue) starting structures. B: The hydroxyl group of Tyr3 points away from the backbone carbonyl of Val21 throughout the 1.2-μs trajectory, because it is relatively tightly packed in a hydrophobic environment. The same is true for Tyr45, which does not flip during the 1.2-μs trajectory. Tyr33, which is more exposed than the other two, undergoes occasional complete ring flips along the trajectory.

To test whether the orientation of the hydroxyl group of Tyr3 is sufficient to bifurcate the dynamical behavior during the trajectory, we rotated the hydroxyl group of Tyr3 by $\sim 180°$ to point towards Val21 in a starting structure derived from the set B simulations (which did not exhibit low order parameters for Asp22) and manually rotated the hydroxyl group of Tyr3 to point away from Val21 towards the solvent for a representative set A starting structure. Eight 2.4 ns NVE simulations were run for each system. Figure 3.11 shows that the behavior is indeed interchanged, demonstrating that the position of the hydroxyl group of Tyr3 at the beginning of the simulation is sufficient to determine the dynamic behavior of Val21/Asp22 during the simulation.

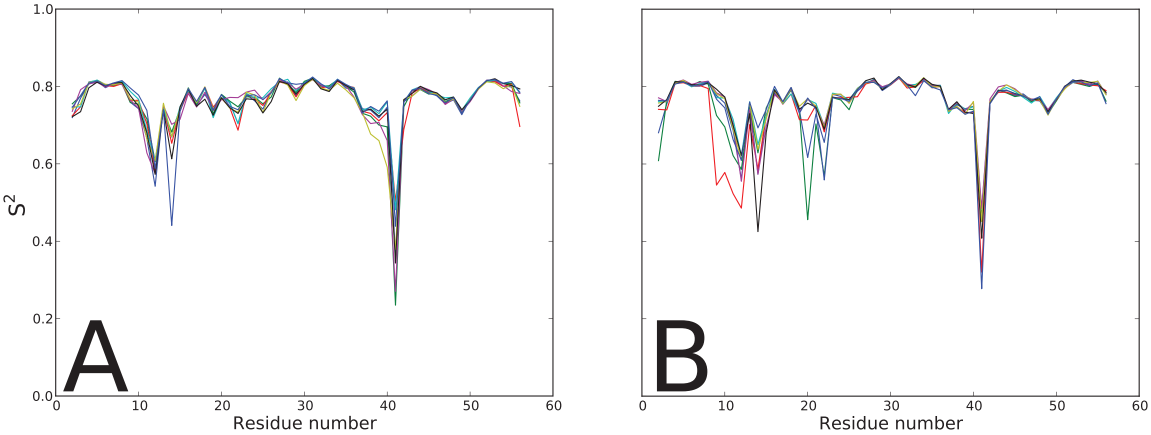


Figure 3.11: **Flipping the Tyrosine 3 OH group changes the behavior of loop 2** After flipping the hydroxyl group of Tyr3 away from Val21 in the set A starting structures (A) or towards Val21 in the set B starting structures (B), the behaviors are reversed. Asp22 now undergoes conformational transitions for the set B but not the set A starting structures, showing that the orientation of the hydroxyl group of Tyr3 is sufficient to determine the dynamical properties of Asp22.

To examine whether the hydroxyl orientation of the set B starting structures (hydroxyl pointing away from loop 2) was so strongly favored that no flip would occur in even a 1.2-μs trajectory, we also ran a 1.2-μs simulation with the same structure used to initiate trajectories shown in Figure 3.11B (set B starting structure with hydroxyl flipped to correspond to set A orientation). Figure 3.12 shows that the χ_1 and χ_2 dihedral angles do not change for Tyr3 throughout both 1.2-μs trajectories, independently of the starting structure. On the other hand, the dihedral angle between the hydroxyl group and the ring changes by 180° at a time point 663 ns into the simulation. Figure 3.13 shows how clearly the hydroxyl flip separates the trajectory into set A like (before the flip) and set B like (after the flip) behavior. A single event does not allow an estimation of the flip rate or populations of the two orientations, but the energy barrier is not too high to be overcome in trajectories on the μs-ms time scale. For all simulations lengths that can easily be achieved with the currently available computational power, the problem remains: the flip rate of the hydroxyl group (much less the entire ring) is too low to equilibrate within nano- or microseconds. The difference in protonation at the beginning still strongly influences the dynamical behavior of the whole trajectory (Figure 3.13).

3.7 Timescales and understanding the discrepancies

Many of the discrepancies between the different simulations and with the experimental data occur for residues that undergo infrequent large transitions in conformational space, leading to the bifurcated behavior described above (Figures 3.4 and 3.6). Many of the trajectories that exhibit exceptionally low order parameters for certain residues also have unconverged autocorrelation functions for those residues (Figure 3.6). Most of those transitions seem to be infrequent enough not to affect the mean order parameters for a very long simulation or a large sample of short simulations that cover many substates in conformational space. These large infrequent transitions might not contribute to NMR spin relaxation rate constants either because signal averaging over the large number of molecules in solution makes them invisible or because the timescales of these motions are beyond those that can be

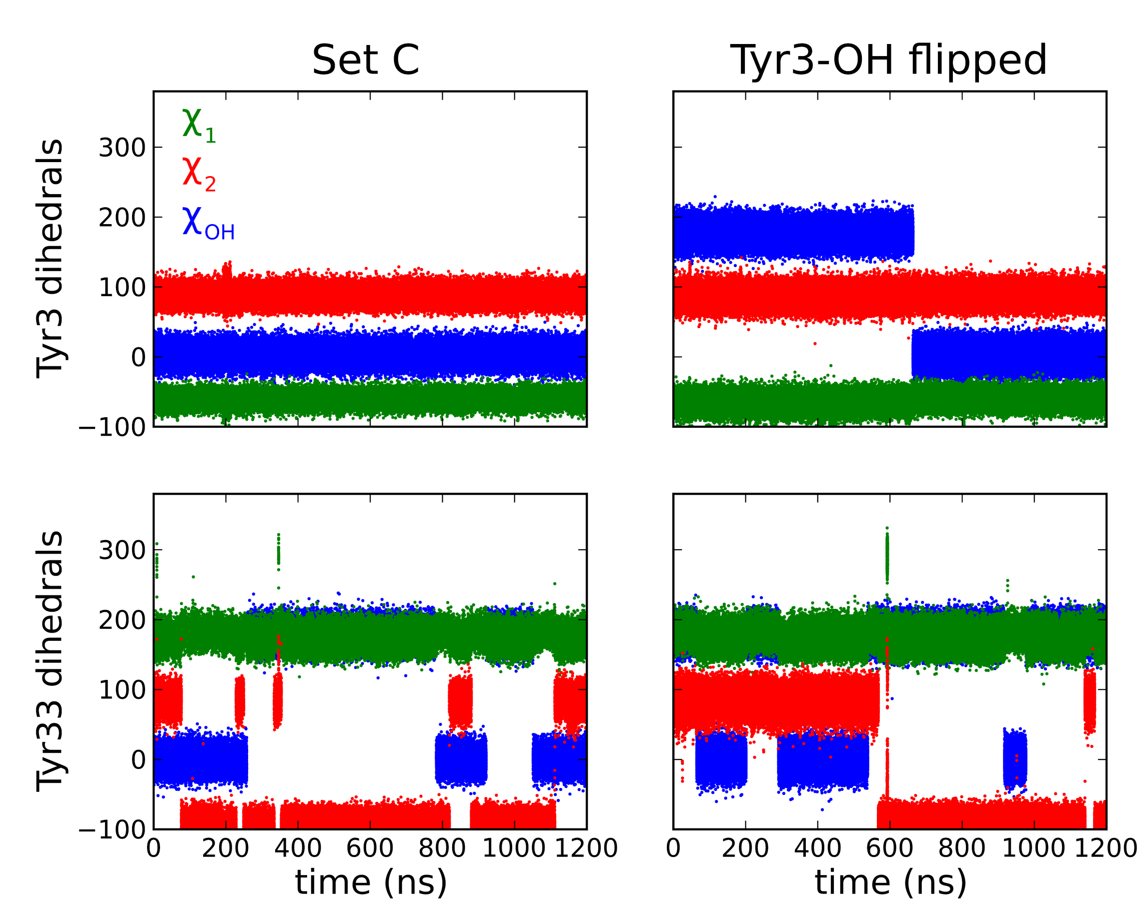


Figure 3.12: **χ-angles of Tyr3 and Tyr33 for the two 1.2-μs trajectories** The original Set C trajectory is shown in the left panel, the simulation started from the structure in which the Tyr3 hydroxyl group was rotated manually to point towards loop 2 is shown in the right panel. χ_1, χ_2 and the dihedral angle towards the hydroxyl group ("χ_{OH}") are represented as green, red and blue dots respectively. In all four cases χ_1 never undergoes a transition. χ_2 and χ_{OH} on the other hand undergo several transitions in Tyr33 on the time scale of the simulations, i.e. the hydroxyl group rotates and the entire ring flip on several occasions, seemingly independently. Tyr 3 never undergoes an entire ring flip in any of the simulations (χ_2) but in the case of the simulation started from the structure with the manually rotated hydroxyl group, it seems to indeed rotate back to its "native" position after approximately 663 ns.

captured by relaxation studies. Therefore, we decided to add a convergence criterion to the autocorrelation functions when averaging over many simulations, as described in the Material and Methods chapter 2. Figure 3.14 shows that including a test of convergence for

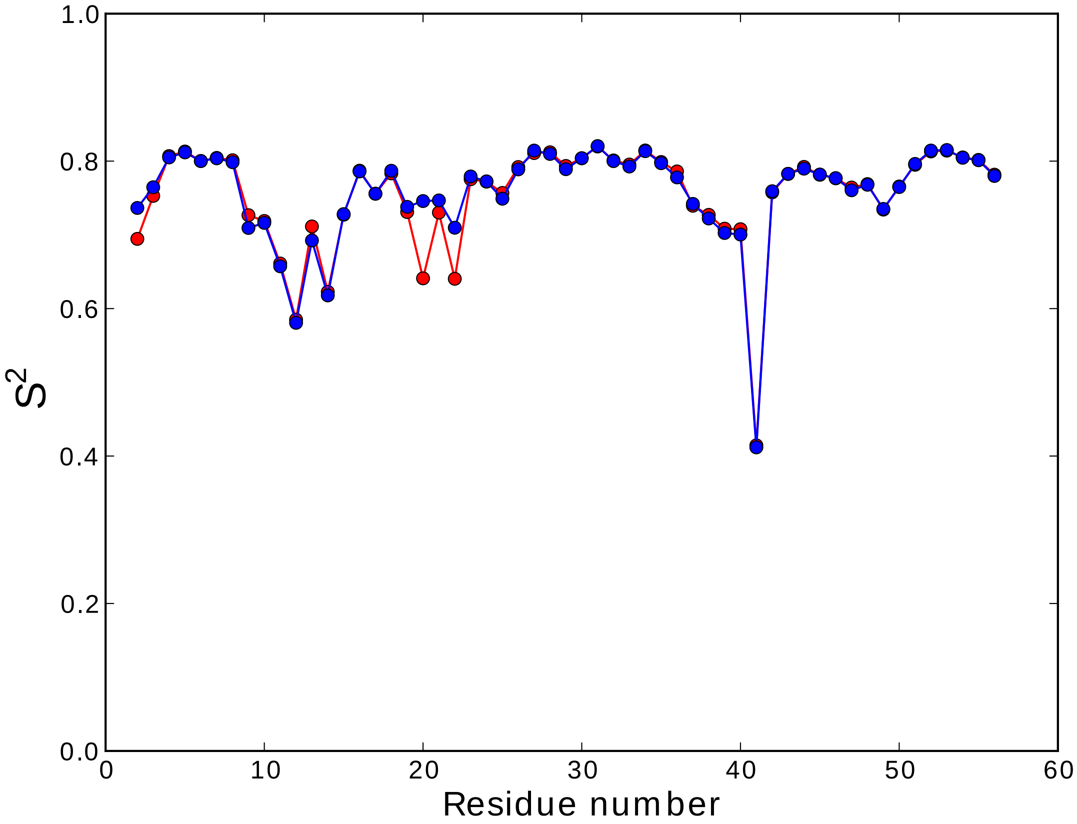


Figure 3.13: **Tyrosine 3 OH flip interchanges behaviors of sets A and B** Order parameters of the 1.2-μs trajectory starting with the Tyr3 hydroxyl group pointing towards loop 2 averaged over all blocks before (red) and after (blue) the hydroxyl flip.

the autocorrelation function increases the order parameters of the main outliers without affecting the remaining residues. This improves the agreement of the different simulations with each other and with NMR-derived order parameters.

3.8 Discussion

An extensive systematic set of simulations of GB3 demonstrates that the choice of the starting structures is more important for the accuracy of the resulting backbone NH order

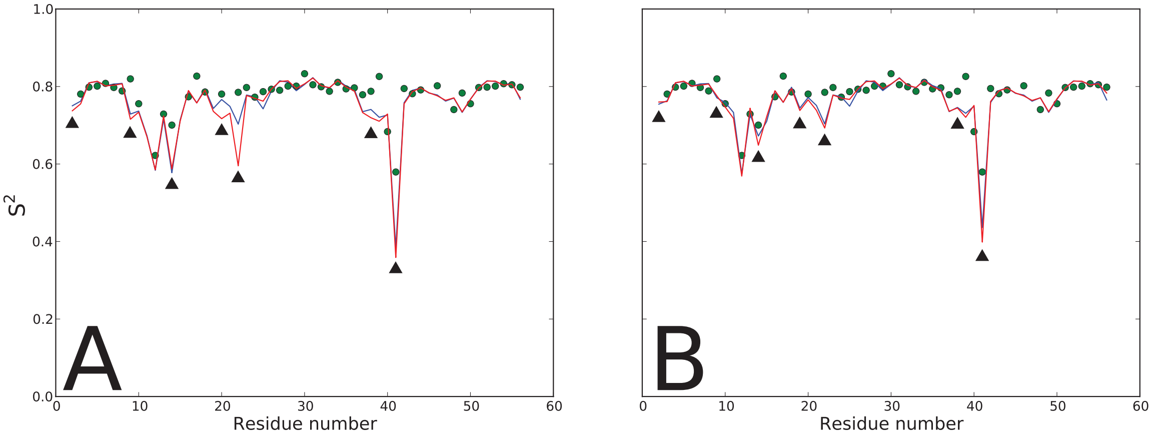


Figure 3.14: **Adding a convergence criterion for autocorrelation functions improves agreement between experiment and simulation**
A: Simulated order parameters for set A (red) and set B (blue) starting structures are shown in comparison to experimentally determined values (green dots). Discrepancies are indicated with black triangles. B: After setting a convergence threshold for the autocorrelation function all the marked regions become more rigid and now agree much better between the two sets of starting structures as well as between simulations and experiment.

parameters than many other variables, including box size or geometry, water model, salt content, force field, or electrostatic cutoff. This result is in agreement with earlier studies on the subject [60, 61]. Additionally, many of the primary outliers of MD simulations in comparison with experimental data also are outliers when comparing the results of different MD simulations, which suggests that sampling is one of the main limitations when comparing NMR- and MD-derived order parameters. A solution state NMR experiment is per se averaging over a large number of molecules in many different states with many different transient and local behaviors. In contrast, each MD simulation considers a single molecule and current computational limitations cannot guarantee ergodicity. Significant improvements have been made to AMBER and CHARMM force fields in recent years [17,58,62,63]. Many of those corrections focus on backbone torsion potentials and have led to improved agreement of simulations with experimental data [41, 64]. Herein we showed that sidechain conformations in starting structures strongly influence backbone order parameters derived from MD simulations and that including recently improved sidechain torsion potentials [17] can shift the simulated populations to a more native area of conformational space independently of the starting structure. The results presented here show that outliers are mostly flexible residues often lying in loops or at the termini of the protein. Those outliers undergo large transitions in conformational space more frequently than other residues. This leads to a bifurcated distribution of their order parameters, reflecting converged and unconverged autocorrelation functions on the timescale of the trajectory or the simulation block used for calculation of the order parameters. The starting structure dictates where the system starts to explore conformational space and consequently how representative sampling will be for the native behavior of the protein. Some transitions were sampled preferentially for one set of starting structures, almost independently of the choice of the parameters used to set up or simulate the system.

Increased simulation lengths allow better sampling, but even for very long simulations the dependence on the starting structure can be strong. The case of Asp22 illustrates this very well. The position of the hydroxyl group of Tyr3 at the beginning of the simulation is sufficient to confine the protein to one part of conformational space for 2.4 ns as well as for 1.2-μs simulations. Earlier studies on the prediction of hydrogen positions have found the accurate prediction of hydroxyl hydrogen positions to be of particular difficulty [65, 66]. Earlier publications have addressed the sampling problem with methods such as accelerated MD (AMD), high temperature MD, replica exchange MD and number of other approaches [67–73]. One of these studies used AMD to generate starting structures for regular MD simulations of GB3 [67]. This approach was not able to produce all of the motions observed in our MD simulations (especially motions in loop 2). On the other hand, any approach that involves the use of a variety of starting structures might mean oversampling experimentally insignificant parts of conformational space.

Here we presented an alternative solution to this problem. The main order parameter outliers between different simulations exhibit bifurcated behavior related to the convergence of the respective bond vector orientational autocorrelation functions. The agreement between the different sets of starting structures as well between simulations and experimentally derived order parameters is improved by excluding simulations that fail a test for convergence of the autocorrelation function of any residue from the averaging of the order parameters for that residue. The success of that strategy might indicate that the unconverged movements are too rare to be seen experimentally in the bulk of molecules, occur on timescales inaccessible to NMR spin relaxation experiments or are erroneously sampled by the simulation and do not occur in the real protein in solution. Thus, that a motion is observed in an MD simulation but not in a specific NMR experiment does not necessarily mean that the force field is erroneous: the motion simply may not be visible with the specific experimental method.

The difference of order parameters resulting from different magnetic field strengths, different chemical shift anisotropies, or methods used for deriving the order parameters has been shown to exceed 0.1 for some residues of GB3, which is similar to differences between simulations in different force fields and between simulations and experimental data [41–43]. More recently, Yao et al. have used site specific ^{15}N chemical shift anisotropy tensors to improve the calculation of order parameters from NMR experiments [43]. Indeed, some areas of the protein, especially the α-helix, now show a much better agreement between experimental and MD-derived order parameters (Figure 3.15) [42,43]. This again illustrates that improvements in experimental methods and interpretation of experimental results also are critical for assessing necessary improvements in simulation methods.

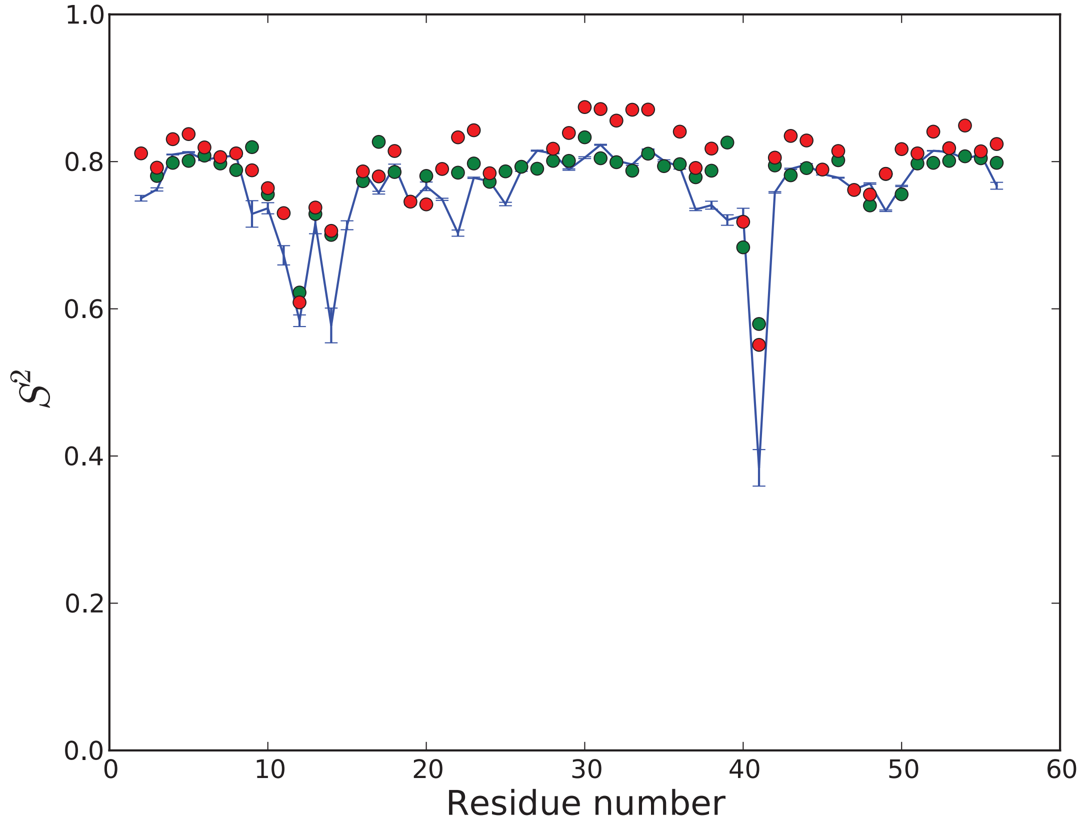


Figure 3.15: **The use of site-specific CSA tensors improves the agreement with MD simulations**

Comparison of the initial set B simulations (blue line) with NMR spin relaxation derived order parameters using site-specific CSA tensors (green circles [43]) and non-site-specific CSA tensors [42]. Improvement using site-specific CSA tensors is particularly striking in the alpha helical region (residues 22-37). Error bars represent standard errors.

Chapter 4

Thermostability of enzymes from molecular dynamics simulations

This chapter is adapted from a manuscript in preparation for publication.

Simulations have been prepared and conducted by Kate Stafford, PhD, as part of her own PhD thesis [35].

4.1 Introduction

Understanding protein stability, and more specifically thermostability, has long been of interest in structural biology and biophysics, but also of biotechnology [74]. Importantly, thermostable enzymes may be useful catalysts for industrial processes run at relatively high temperatures. The use of higher temperatures has many advantages, including increased reaction rate, increased solubility of reactants, and reduced contaminating microbial growth [74–76].

The unfolding, or melting temperature, T_m is a metric for the thermostability of a protein. T_m is the temperature at which the Gibbs free energy ΔG of the folded and unfolded state, and thus their populations, are equal. While attempts have been made to determine T_m by molecular dynamics simulations, computational limitations usually limit the study of complete unfolding processes to very small and fast folding proteins [77,78]. Several computational tools to predict protein stability have been developed falling into two main categories. The first studies the energy of unfolding (ΔG) of proteins by using physical, statistical or empirical potentials, and the second by using machine-learning methods trained on datasets of experimental unfolding energies. Combined approaches also exist. Assessments of several of these approaches can be found in references [79–81]. Promising results have been reported recently by using Monte Carlo simulation approach to the bacterial enzyme dihydrofolate reductase [82]. A more high-throughput approach has been developed for finding multiple stabilizing mutations of a protein [83].

Ribonuclease HI (RNase H; EC 3.1.26.4) enzymes, have been studied extensively to shed light on thermostability, because structurally highly conserved homologs exist in both psychrotrophic and thermophilic organisms [38,39,84–94]. These enzymes non-specifically cleave the RNA strand of RNA:DNA hybrid substrates [95], and have been implicated in many biological processes, including removal of R-loops, removal of Okazaki fragments, synthesis of multicopy single-stranded DNA, and removal of ribonucleotides misincorporated into the genome [38,96].

4.2 Correlation of temperature dependence of simulated order parameters with experimentally determined melting temperatures

To study the influence of temperature on the dynamical behavior of RNase H variants *in silico*, we conducted Molecular Dynamics (MD) simulations for RNase H enzymes from

Table 4.1: **Ribonuclease H enzymes.**

Protein	PDB ID	Source Organism	T_m ($°C$)
soRNH	2E4L	*Shewanella oneidensis*	53.2 [97]
ctRNH	3H08	*Chlorobium tepidum*	68.5 [98]
ecRNH	2RN2	*Escherichia coli*	70.7 [97]
ttRNH	1RIL	*Thermus thermophilus*	89 [93]

four different organisms, as shown in Table 4.1 [38]. The set is composed of proteins from psychrotrophic (soRNH), mesophilic (ecRNH), moderately thermophilic (ctRNH), and thermophilic (ttRNH), organisms. Despite its origins in the proteome of a moderate thermophile, the protein ctRNH has a melting temperature T_m that is similar to that of the mesophile protein ecRNH.

The four proteins were simulated for 100 ns at 273, 300 and 340 K. The preparation of the simulations has been conducted by Dr. Kate Stafford and has been described previously [35–39]. All 12 trajectories were divided into 10 ns blocks to reflect global tumbling time [39, 99] and the square of the generalized order parameters (S^2) for the backbone NH bond vector (henceforth simply "order parameters") were calculated and averaged over all blocks for each trajectory. The order parameters describe the orientational fluctuations of the backbone NH vectors and can by measured experimentally by NMR spectroscopy [34]. Order parameters for all proteins at the three studied temperatures are shown in Figure 4.1. All order parameters were scaled by $\xi = (1.02/1.04)^6 \approx 0.89$ to account for zero point vibrational motions of the NH bond vectors [34, 54].

The temperature dependence of S^2 is described by the dimensionless parameter Λ, which is directly related to the temperature dependent effective potential for the fluctuations of the NH bond vector, and together with S^2 provides information about the contributions of

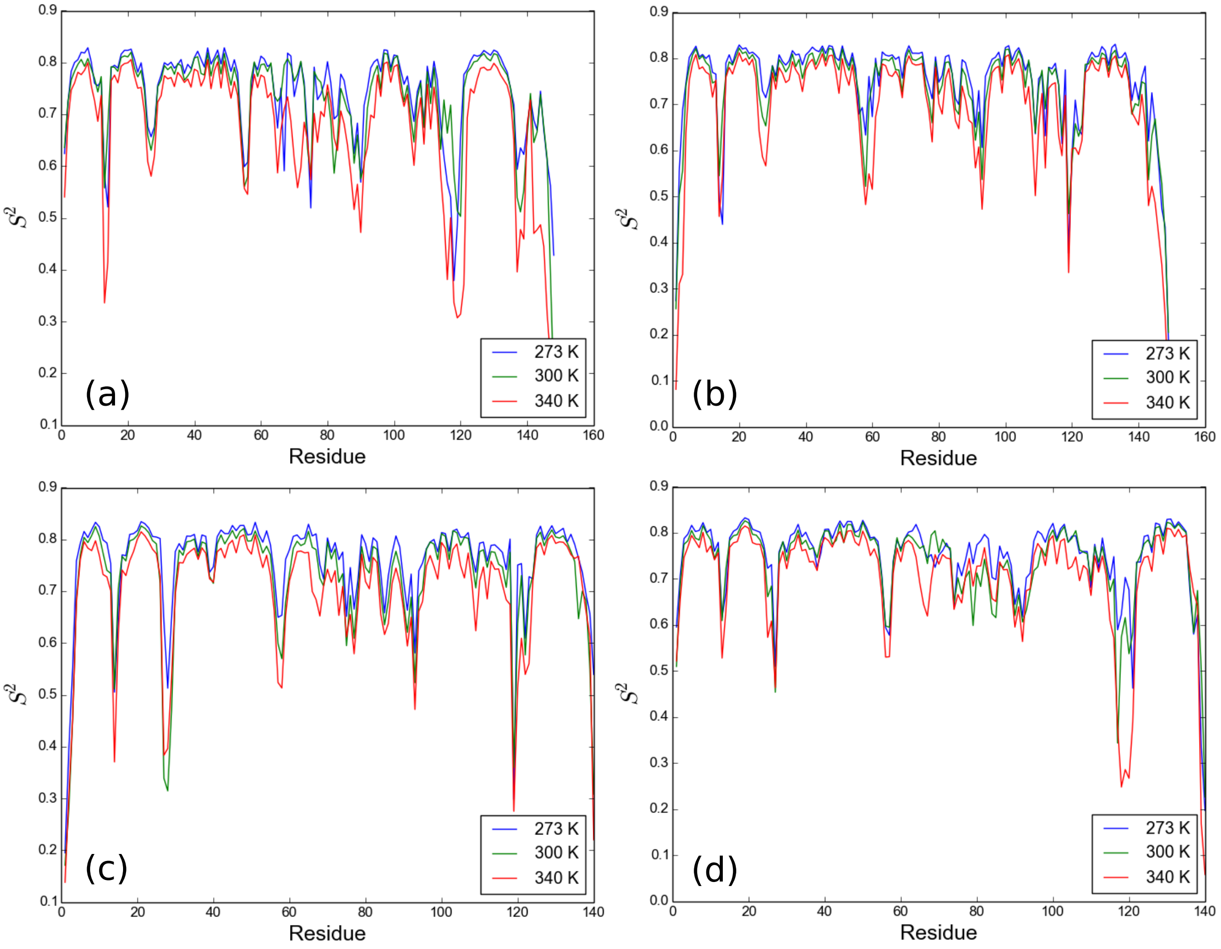


Figure 4.1: **Backbone ^{15}N squares of the generalized order parameters for all four organisms**
Backbone ^{15}N squares of the generalized order parameters for (a) *S. oneidensis* RNase H at 273, 300 and 340 K, (b) *E. coli* RNase H at 273, 300 and 340 K, (c) *C. tepidium* RNase H at 273, 300 and 340 K and (d) *T. thermophilus* RNase H at 273, 300 and 340 K.

these fluctuations to heat capacity [100–102]:

$$\Lambda = \frac{dln(1-S)}{dlnT} \tag{4.1}$$

Figure 4.2 shows plots of $ln(1-S)$ against lnT for select residues in *E. coli* RNase H. Linear regression was used to extract Λ values as the slope of these plots for each residue. The quality of the fit was assessed by the correlation coefficient R. Incorporating simulations at additional temperatures has only minor effects on resulting Λ values (Figure 4.3).

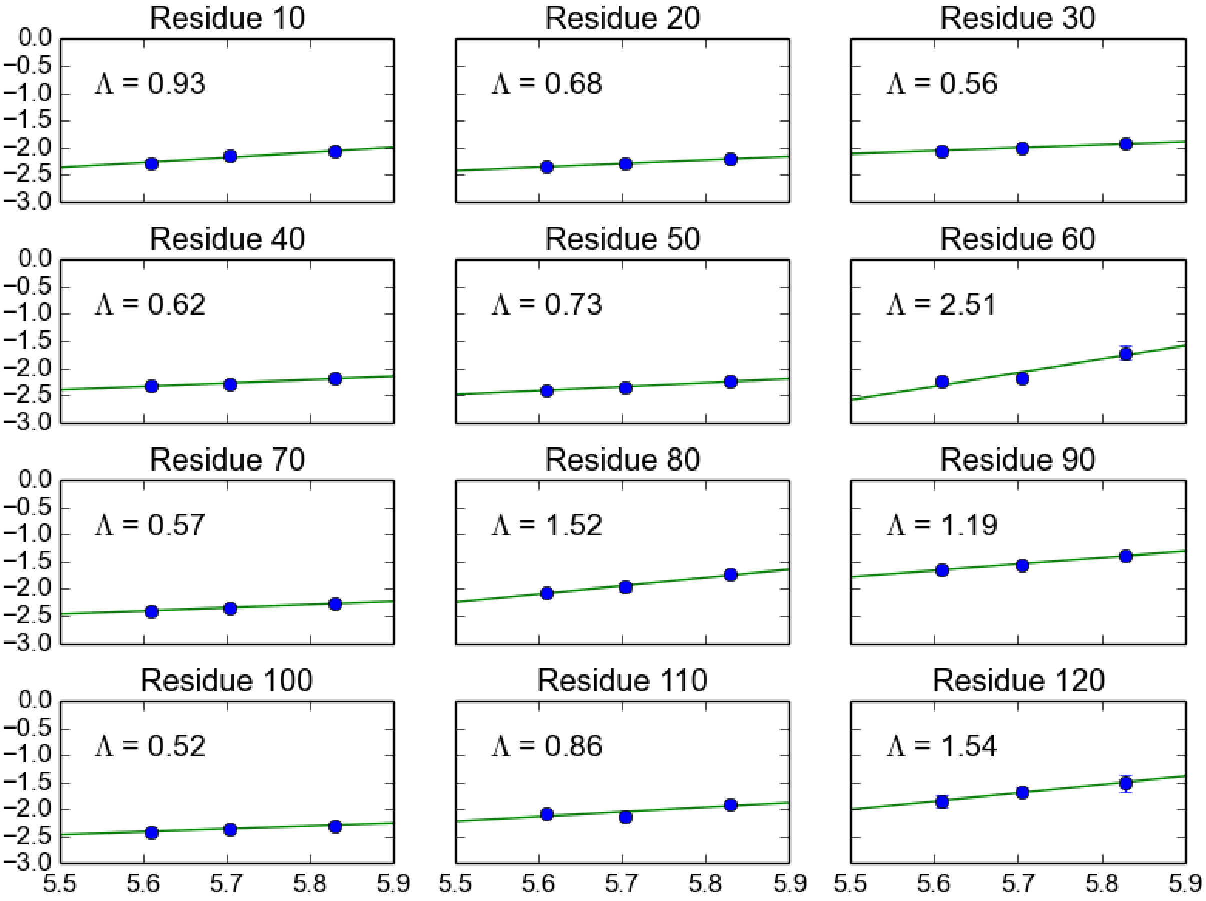


Figure 4.2: **Log Plots for *E. coli* at three temperatures**
Plots of $ln(1-S)$ vs. lnT for *E. coli* RNase H at three different temperatures: 273 K, 300 K, 340 K. The parameter Λ corresponds to the slopes of the green lines, which are obtained by linear regression.

Figure 4.4 shows Λ and R values as a function of sequence for the four studied proteins. Because Prolines do not have an NH bond vector, they are omitted from the sequences for this figure. For most residues the linear regression yielded R values close to unity. The majority of the low R values correspond to low Λ values and thus low temperature

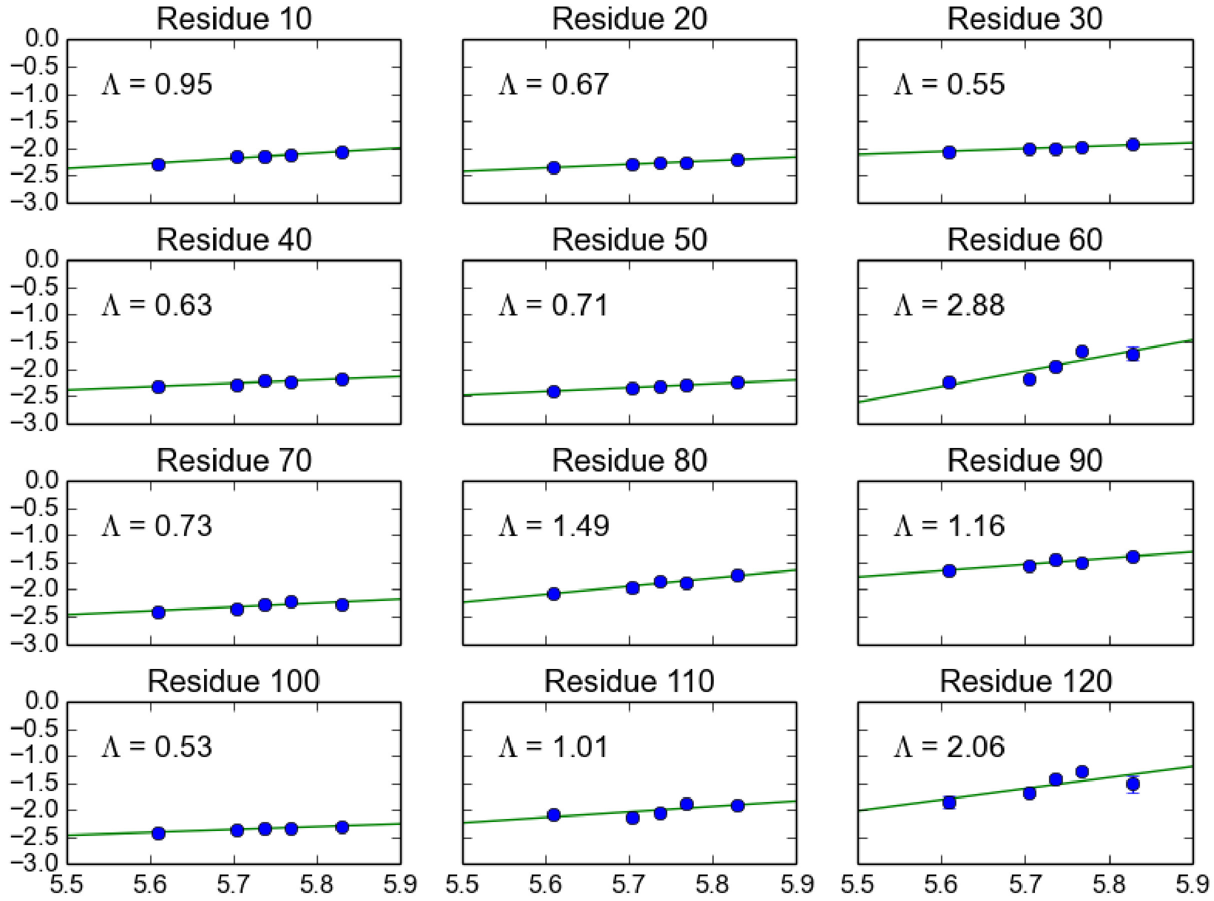


Figure 4.3: **Log Plots for E. coli at five temperatures**
Plots of $ln(1-S)$ vs. lnT for *E. coli* RNase H at five different temperatures: 273 K, 300 K, 310 K, 320 K, 340 K. The parameter Λ corresponds to the slopes of the green lines, which are obtained by linear regression.

dependence. Accordingly, the number of residues with R values smaller than 0.8 is much smaller for ecRNH (15 out of 149) than for ttRNH (38 out of 140). The distributions of Λ for the different proteins are compared in Figure 4.5.

A plot of Λ averaged over all residues for each organism against experimental melting temperature Tm is reveals a homomorphic relationship between the two, with a lower average Λ value corresponding to a higher melting temperature (Figure 4.5). A similar correlation is observed between Λ and ΔG for folding calculated at 304 K, the midpoint of the investigated temperature range, using the Gibbs-Helmholtz equation (not shown).

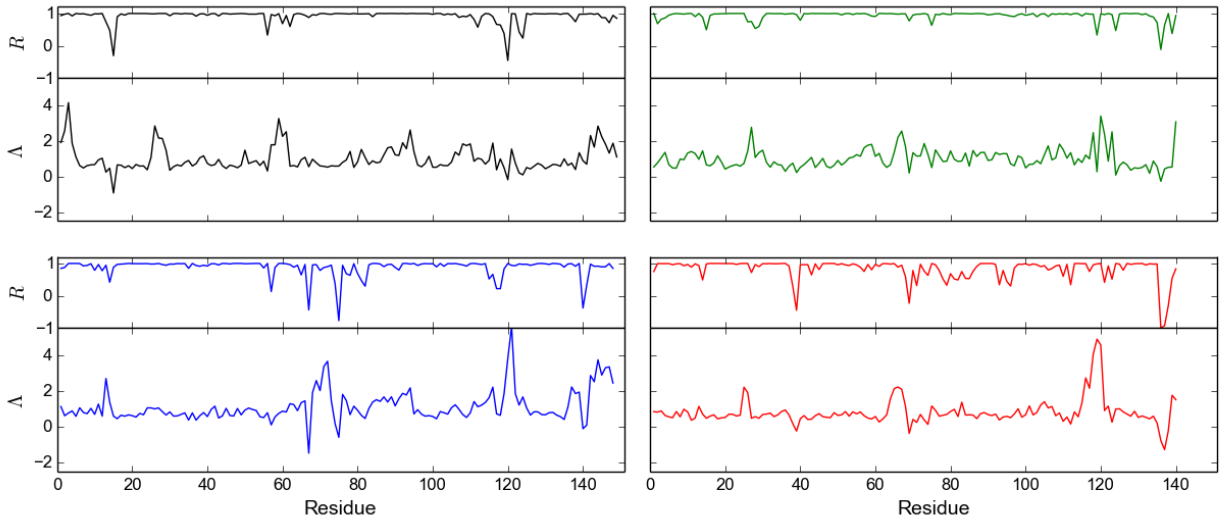


Figure 4.4: **Λ and R values for all four organisms**
Λ and R values for ecRNH (black), ctRNH (green), soRNH (blue), ttRNH (red).

The linear relationship of all four data points is statistically highly significant ($p < 0.01$ for weighted least squares). Removing Λ values of residues with low R values increases the average Λ values in a similar fashion for all proteins, without affecting the qualitative relationship between average Λ and T_m, the slope of the fitted line is −0.008 or −0.009 with R values of 0.98 or 0.99, if these residues are excluded or included, respectively.

4.3 The point mutant iG80b

To study the effect of a single thermostabilizing mutation on the Λ values of a protein, we conducted another set of MD simulations on an E. coli variant with a thermostabilizing Glycine insertion at position 80b (mimicking this position in ttRNH); the simulated order parameters are shown in Figure 4.6a. This mutant has been crystallized (PDB: 1GOA) and thermodynamically characterized [103, 104]. The insertion has a small thermostabilizing

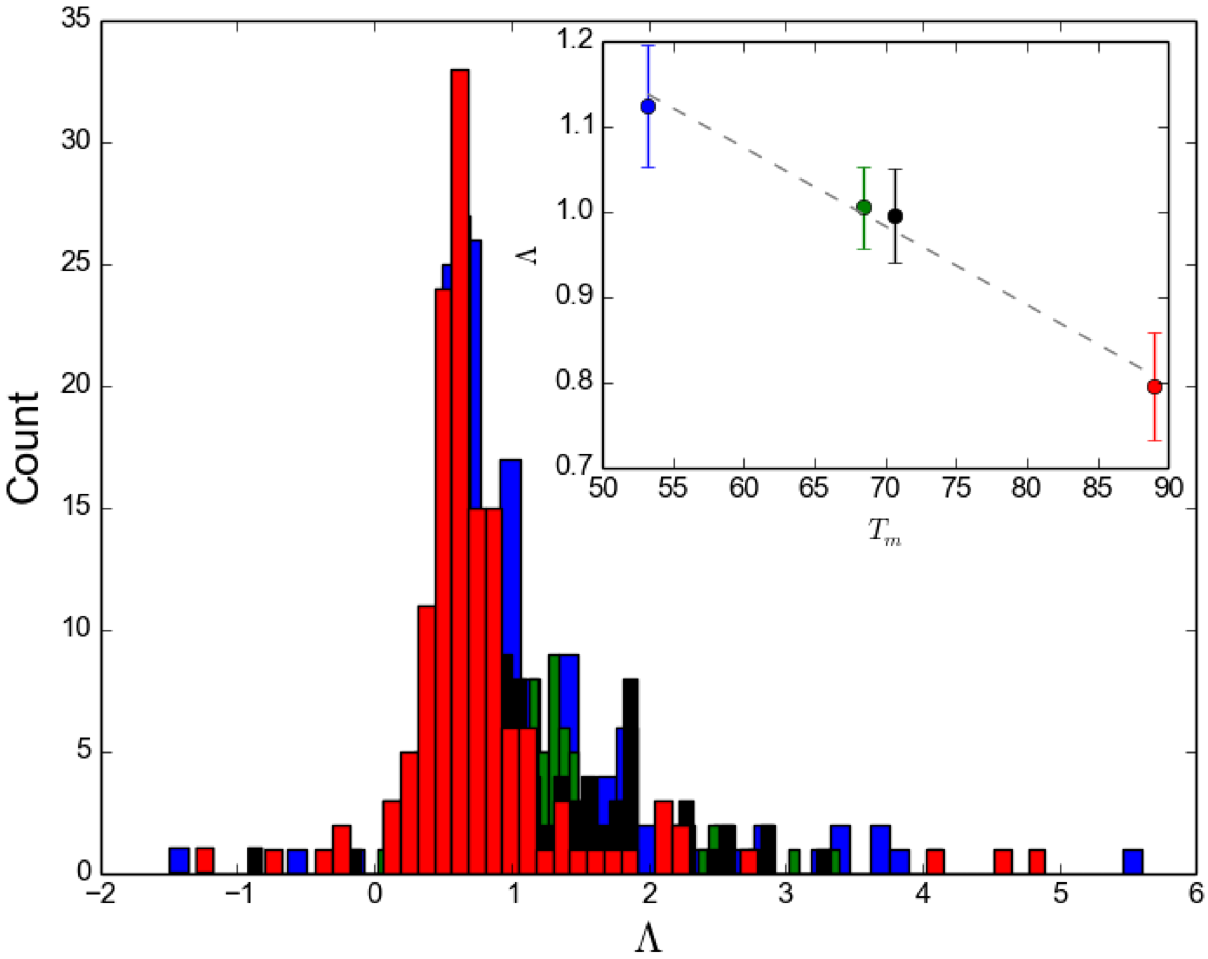


Figure 4.5: **Average Λ values reveal a homomorphic relationship with melting temperature**
Histograms of Λ values for the four organisms with experimentally determined T_m values: soRNH (blue), ctRNH (green), ecRNH (black), ttRNH (red). The inset shows the homomorphic relationship between average Λ values and experimentally determined T_m.

effect and increases T_m by 1.2 K [103, 104]. Figure 4.6b and 4.6c show that the correlation between Λ values for ecRNH and the iG80b mutant are extremely high for about half of the residues, while the rest differ quite significantly. The insertion seems thus to have effects that are not restricted to the proximity of the insertion but lead to changes across the protein, presumably through changes in packing (ttRNH has a reduced solvent accessible surface area compared to ecRNH - ca. 8800Å^2 vs ca. 9500Å^2 - , the iG80b mutant lies in between with ca. 9200Å^2). The average value for Λ decreases slightly from 0.995 to 0.965, agreeing with the increase in T_m of 1.2 K. Figure 4.6d shows that including this fifth datapoint improves the statistical significance of the weighted linear regression further (p $<$ 0.001).

4.4 Regions important for thermostability *in silico* correspond to regions important for substrate binding and thermostability *in vitro*

A plot of the Λ values of the four wild-type proteins and the iG80b mutant, accounting for insertions and deletions, as well as the positions of prolines, (Figure 4.7) shows regions with strong differences in Λ values especially for the psychotrophic (soRNH) and thermophilic (ttRNH) proteins. These regions include helix α_B, the handle region and the β_5/α_E-loop, which have all been implicated in substrate binding [39, 84, 105]. Figure 4.8 shows the structures of soRNH and ttRNH with residues with particularly high or low Λ values highlighted. Many of these residues have been implicated in thermostability as well [105, 106]. As one example, the octapeptide LKKAFTEG in helix α_B of ttRNH (comprising the iG80b insertion) has an average Λ value of 0.4±0.2 in ttRNH and an average of 2.2±0.4 for the corresponding heptapeptide MRQGIMT in soRNH. This ttRNH octapeptide corresponds to the region R5 described in [106], which when replacing the corresponding region in ecRNH leads to a $\sim$5 K increase in T_m at pH 5.5. The observation that many of these residues are important for both substrate binding and thermostability may reflect evolutionary trade-offs between activity and stability.

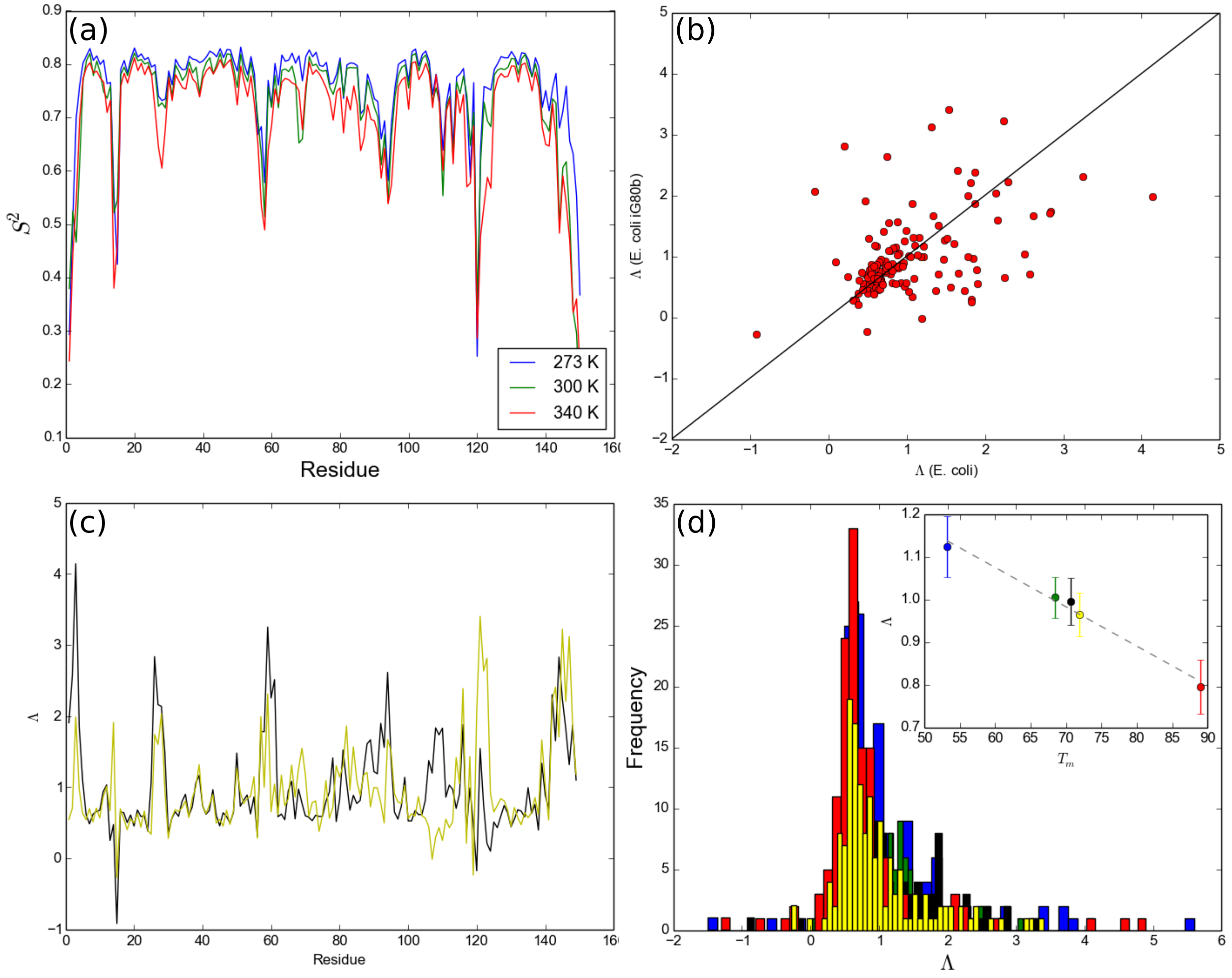


Figure 4.6: **Backbone ^{15}N squares of the generalized order parameters for the iG80b mutant**

(a) Backbone ^{15}N squares of the generalized order parameters for the iG80b insertion *E. coli* RNase H mutant. (b) Correlation plot of Λ values between *E. coli* RNH and the iG80b mutant. (c) Λ values for *E. coli* RNase H in black and the iG80b mutant in yellow. (d) This figure corresponds to Figure 4.5, but with the E. coli iG80b mutant added in yellow. Histrograms of Λ values for the five organisms with experimentally determined T_m values: soRNH (blue), ctRNH (green), ecRNH (black), ttRNH (red) and ecRNH iG80b (yellow). The inset shows the homomorphic relationship between average Λ values and experimentally determined T_m.

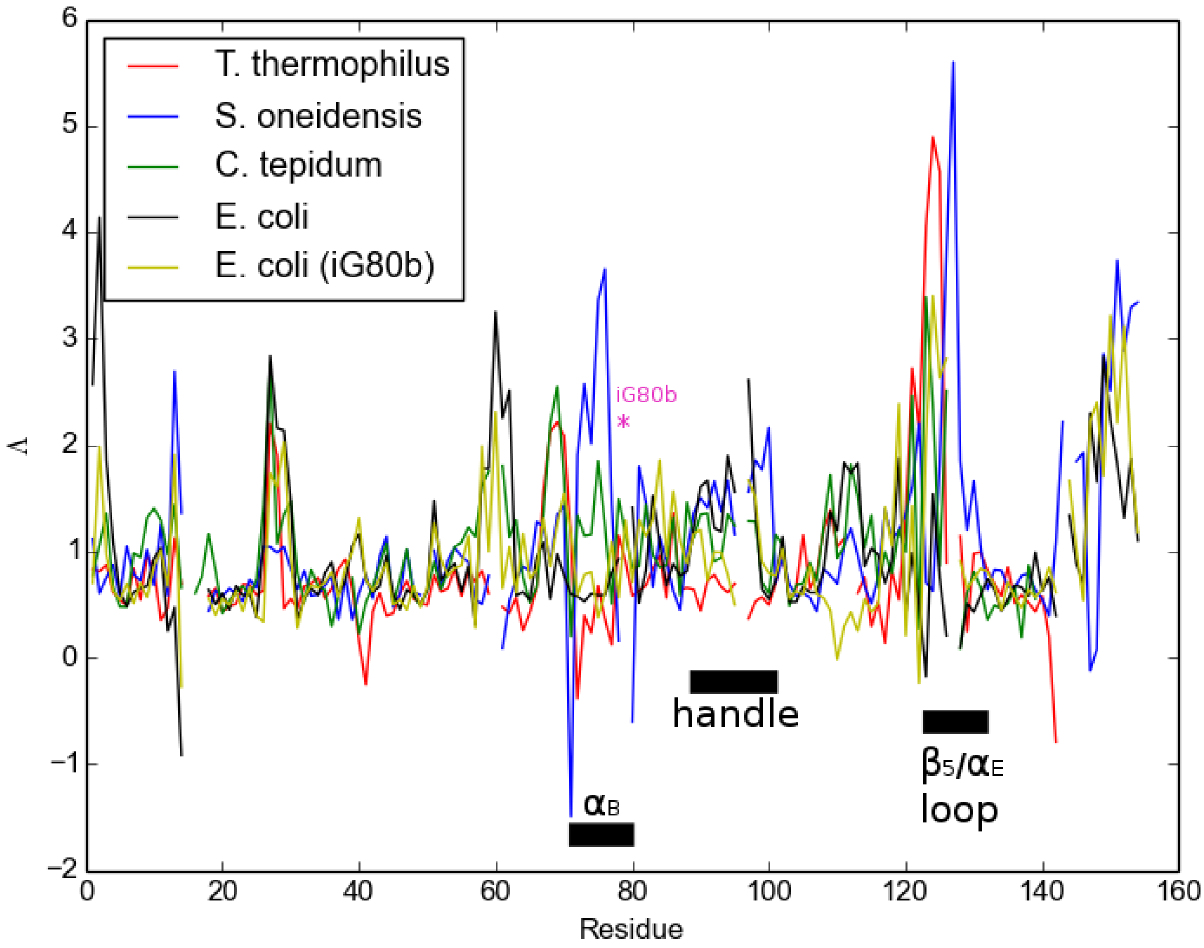


Figure 4.7: Λ **values for the five organisms by sequence alignment** Missing values are either insertions/deletions or Prolines. The Glycine insertion is indicated by a pink asterisk and three regions with large differences between the psychotroph soRNH and the thermophile ttRNH are indicated by black bars.

4.5 Discussion

We have shown, at least within one homologous family of enzymes including proteins from psychrotrophic, mesophilic and thermophilic organisms, that a linear homomorphic relationship exists between experimentally determined melting temperature and the average of the simulated parameter Λ, describing the temperature dependence of conformational fluctuations of the backbone amide bond vectors. This result suggests that Λ is a good proxy

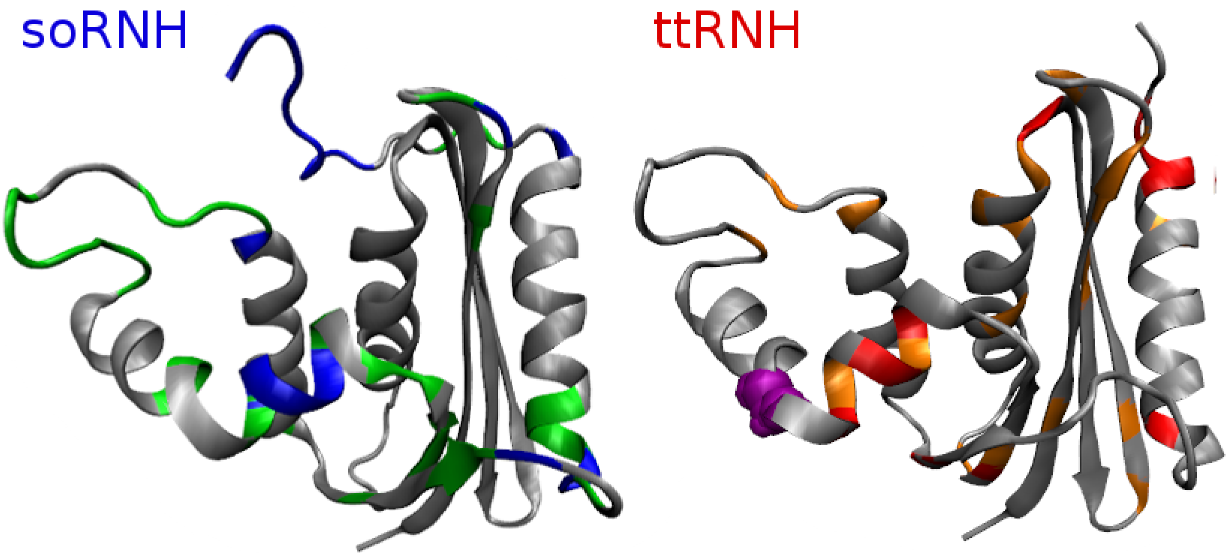


Figure 4.8: **Residues with strong influence on average Λ correlate with residues important for *in vitro* thermostability and substrate binding**
Structures of soRNH (left) and ttRNH (right). soRNH: Residues with Λ values over 1.2 in green, residues with Λ values over 2 in blue. ttRNH: Residues with Λ values under 0.5 in orange, residues with Λ values under 0.3 in red. The iG80b in ttRNH insertion is shown as purple spheres.

for thermostability within a family of homologous proteins. Lastly, this study suggests the possibility for *in silico* identification of thermostabilizing mutations within a protein family, without the need to simulate full unfolding events, as an adjunct to other computational approaches for protein design, potentially in combination with more high-throughput pipelines for protein stabilization [83].

Chapter 5

Study of an Scr-Cofactor-DNA complex by NMR spectroscopy and MD simulations

5.1 Introduction

5.1.1 The Hox cofactors Homothorax and Extradenticle

Hox proteins are transcription factors that function as regulators of development and are conserved across many species from fruit flies to humans [107]. They define the cell fate along the antero-posterior (AP) axis of bilaterian organisms and specify segment identity during early embryonic development. They are encoded by the Hox genes, which are named after a shared circa 180 nucleotide sequence called homeobox. The homeobox translates into a 60 amino acid helix-turn-helix structure called homeodomain (HD), which can bind to DNA (Figure 5.1).

Transcription factors must select a subset of DNA sequences out of a very large number

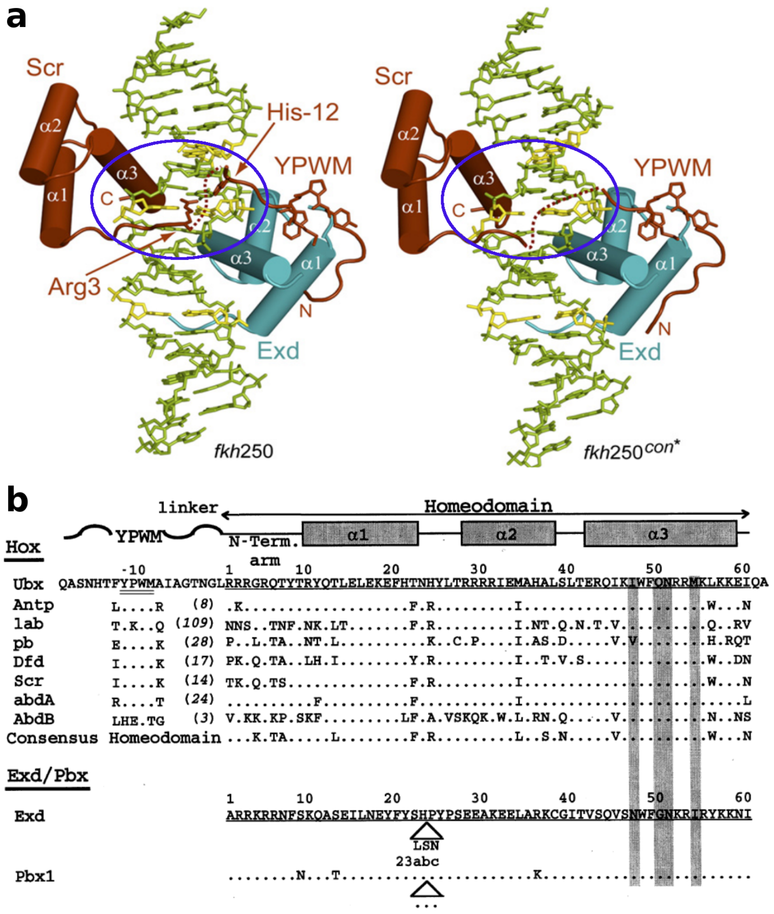

a
Scr
α2
α1
α3
C
His-12
YPWM
α2
α3
α1
Arg3
N
Exd
fkh250
Scr
α2
α1
α3
C
YPWM
α2
α3
α1
N
Exd
fkh250con*
b
linker
Homeodomain
YPWM
N-Term. arm
α1
α2
α3
Hox
-10 1 10 20 30 40 50 60
Ubx QASNHTFYPWMAIAGTNGLRRRGRQTYTRYQTLELEKEFHTNHYLTRRRRIEMAHALSLTERQIKIWFQNRRMKLKKEIQA
Antp L....R (8) .K..................F.R.........I.....................W...N
lab T.K..Q (109) NNS..TNF.NK.LT.......F.R.........I.NT.Q.N.T.V...........Q..RV
pb E....K (28) P..L.TA..NT.L..........K..C.P....I.AS.D.....V.V..........H.RQT
Dfd I....K (17) PK.Q.TA..LH.I........Y.R.........I..T.V.S...............W..DN
Scr I....K (14) TK.Q.TS..............F.R.........I.......................W...N
abdA R....T (24) .........F..........F...........I.............................L
AbdB LHE.TG (3) V.KK.KP.SKF.........LF.A.VSKQK.W.L.RN.Q.....V...........N..NS
Consensus Homeodomain ...K.TA.....L........F.R.........L..S.N.....V...........W...N
Exd/Pbx
1 10 20 30 40 50 60
Exd ARRKRRNFSKQASEILNEYFYSHPYPSEEAKEELARKCGITVSQVSNWFGNKRIRYKKNI
LSN
23abc
Pbx1 ........N...T.......................K.........................
...

Figure 5.1: **Structure of Scr and Exd bound to DNA and sequence alignment of Drosophila Hox proteins**

a) Crystal structures of Scr and Exd homeodomains bound to a specific (fkh250, left) and generic (fkh250CON, right) sequence. The NTA and linker region are much more resolved in complex with the specific sequence (circled in blue). Reprinted (adapted) from Cell, 131, Joshi R, Passner JM, Rohs R, Jain R, Sosinsky A, et al., Functional specificity of a Hox protein mediated by the recognition of minor groove structure. 530 - 543, Copyright 2007, with permission from Elsevier. [49] b) Sequence alignment of different *Drosophila* Hox homeodomains with linker regions in reference to Ubx. Secondary structures are shown on the top. The four DNA contacting residues in the recognition helix are shaded in gray. The linker region is not considered part of the HD and thus numbered negatively. The numbers in parentheses represent the distances of the YPWM motif from the HD (in number of residues). Reprinted by permission from Macmillan Publishers Ltd: Nature, 397, Passner JM, Ryoo HD, Shen L, Mann RS, Aggarwal AK. Structure of a DNA bound Ultrabithorax-Extradenticle homeodomain complex. 714 - 719, copyright 1999. [108]

of potential binding sites that are typically present in eukaryotic genomes. Hox proteins regulate a large number of genes in flies, humans and other animals and many of the target genes are highly specific for a certain Hox paralog [109]. For instance, the Hox protein Scr in Drosophila for example is the only paralog that can initiate salivary gland formation [110]. On the other hand, some functions can be carried out by more than one Hox protein. For instance, many Hox proteins in Drosophila, including Scr, can repress the antennal-specifying gene homothorax (hth) [111]. This suggests that some binding sites are paralog specific while others have a lesser degree of specificity or lack paralog specificity altogether.

All Hox proteins share the conserved HD domain and they all bind to very similar DNA sequences [112]. The third helix of the HD ("recognition helix") makes nearly identical contacts to the bases in the major groove of the DNA in all available structures, and cannot account for the *in vivo* specificity of the different Hox proteins. Many Hox binding sites have been shown to require cofactors for binding, such as the proteins Extradenticle (Exd, Pbx in vertebrates) or Homothorax (Hth, Meis in vertebrates) [113, 114]. The initiation of salivary gland formation by Scr for example is driven by an Scr-Exd-Hth complex activating the target site fork head (fkh) [115].

Several studies suggest that the HD N-terminal arm (NTA) is crucial for specificity of Hox proteins (reviewed in [116]). The NTA is disordered in most available NMR and crystal structures including two structures of a Hox-Exd/Pbx-DNA ternary complex [108, 117]. The work of Joshi et al. [49] suggests that this disorder is related to the use of high-affinity consensus binding sequences instead of high-specificity *in vivo* binding sequences. They solved crystal structures of a ternary Scr-Exd-DNA ternary complex with specific (fkh250) and non-specific (fkh250CON) DNA binding sites, showing that the specific interaction of Scr with fkh250 seems to be mediated by the NTA and linker region contacting the minor groove of the DNA (Figure 5.1). This region contains Histidine and Arginine residues, which are conserved among Scr orthologs and insert into the minor groove of a specific binding sequence but which are disordered in the crystal structure with fkh250CON. The N-terminus of Scr also interacts with Exd, which might thus function to properly position the NTA and linker for specific recognition of the minor groove. The group proposed that Hox proteins recognize "generic" binding sites through the interaction of their conserved HDs with the major groove of the DNA, but select specific sites with their NTA by shape-recognition of the DNA with the help of cofactors such as Exd.

Due to the absence of electron density for the N-terminal loop region in most crystal structures and in particular for Scr bound to both its *in vivo* and a consensus sequence [49], studying its role in DNA recognition is difficult. We were especially interested to see transitions between ordered and disordered states of the N-terminal arm, linker regions and YPWM motif first upon binding of Scr to the DNA and then upon binding of its cofactor Exd. NMR and especially NMR relaxation studies are well suited to study such systems, because they are sensitive to motions on many time scales.

Earlier studies have suggested that Scr, in cooperation with Exd, can recognize preformed minor groove minima of the DNA with the help of the conserved Arginine 3, without addressing why other Hox proteins, which also contain an Arginine at position 3, seem unable to recognize the same sequences. We postulated that this is due to the less highly conserved neighboring residues (in particular residues 4 and 6) influencing the conformational dynamics of the NTA and thus Arginine 3 insertion into the minor groove. The importance of those neighboring residues for shape readout has since been underlined [118].

This chapter describes our attempt to address these questions with NMR spectroscopy and MD simulations. Expression protocols, complex formation conditions and NMR conditions were developed and optimized to allow for the study of the complex. An array of preliminary NMR experiments both on the protein and DNA moieties have been carried out and show cooperative binding of the two homeodomains to the DNA as well as conformational changes in Scr upon binding of its cofactor Exd.

In addition, we conducted MD simulations of Scr as well as mutants at positions 4 and 6 of the N-terminal arm, on the specific and the consensus binding sites fkh and fkhCON to further shine light on the importance of those residues for shape recognition and conformational dynamics of the NTA. Indeed, we were able to find strong correlations of those positions with the conformational dynamics of Arginine 3, further supporting the idea of their role in specific DNA shape readout.

The role of the second Hox cofactor Hth remains elusive. Hth, as well as Exd's PBC domain (interacting with Hth) are poorly understood and structurally not characterized. Here we optimized expression and purification protocols for a large Exd-HM complex containing residues 37 to 311 of the full length Exd protein (including both PBC domains) and a large part of Hth's HM domain (residues 79 to 247). We also conducted preliminary NMR relaxation experiments of this complex alone and together with DNA. These experiments suggest independent tumbling of the PBC and HM domains from the Exd homeodomain, suggesting that they can be expressed and characterized independently from the homeodomain. This work is lays the groundwork for and encourages future studies of a Hox-DNA-HM/Exd complex by NMR spectroscopy and X-ray crystallography.

5.1.2 Previous attempts at studying the complex

Early attempts to study Exd and Hth with NMR spectroscopy were conducted by Keri Siggers in the context of her PhD thesis in the Palmer laboratory [22]. She started by trying to express full-length Exd and Hth but observed extensive aggregation for Exd and degradation for Hth. She then tried to work only with the proposed interaction domains of the two proteins. Exd 1-126 (comprising the PBC-A domain of Exd) still aggregated and the HM domain (Hth 91-209) alone was insoluble. Using secondary structure predictions she tried to further optimize the constructs to Hth 70/76-213, Exd 29-122/144 but again the proteins turned out to be insoluble. Refolding and coeluting the two proteins still led to aggregation.

The first step towards soluble complex was a coexpression of both full-length proteins. This seemed to yield acceptable amounts of soluble protein that did not aggregate on a native gel. The resulting complex underwent partial proteolysis using a protease cocktail of Papain, V8 and Trypsin, followed by a western blot and N-terminal sequencing to identify minimal soluble interaction domains (the sensitivity of Hth to proteolysis made the results difficult to interpret however). Two resulting constructs seemed to be soluble, namely HM 1-225/FL-Exd and HM 85-209/Exd 36-236. It should be noted that the latter construct of Exd is missing the homeodomain, which is essential for interaction with DNA and Hox.

The smaller complex (HM 85-209/Exd 36-236, HD-less) showed very few peaks in an HSQC spectrum, meaning that the complex is either still aggregating or largely disordered.

The next attempt to study the complex by NMR was performed by Nichole O'Connell in the context of her PhD thesis in the Palmer laboratory [23]. Her strategy was to look at only the homeodomains of Scr and Exd and their interaction with a specific *in vivo* DNA target sequence (fkh) and a derived consensus sequence (fkhCON), such as described by Joshi et al. [49]. The constructs she used were Scr 298-285 C362S and Exd 238-320 (both HD only). She was able to assign the backbone resonances of 62 out of the 88 Scr residues, including a large part of the NTA (such as the RQR motif) and linker regions.

Unfortunately, concentrations of soluble complex were low because of precipitation upon titration of DNA and protein. A solubility screen yielded a small number of conditions leading to little or no precipitation but these conditions turned out to show no binding in NMR experiments. The conditions used for the set of NMR experiments that showed binding were: 20mM $NaPO_4$ pH 7.0 or pH 8.0, 50mM NaCl, 5mM TCEP, 10% D2O, 0.02% NaN_3. All other tested conditions showed either no binding or excessive precipitation.

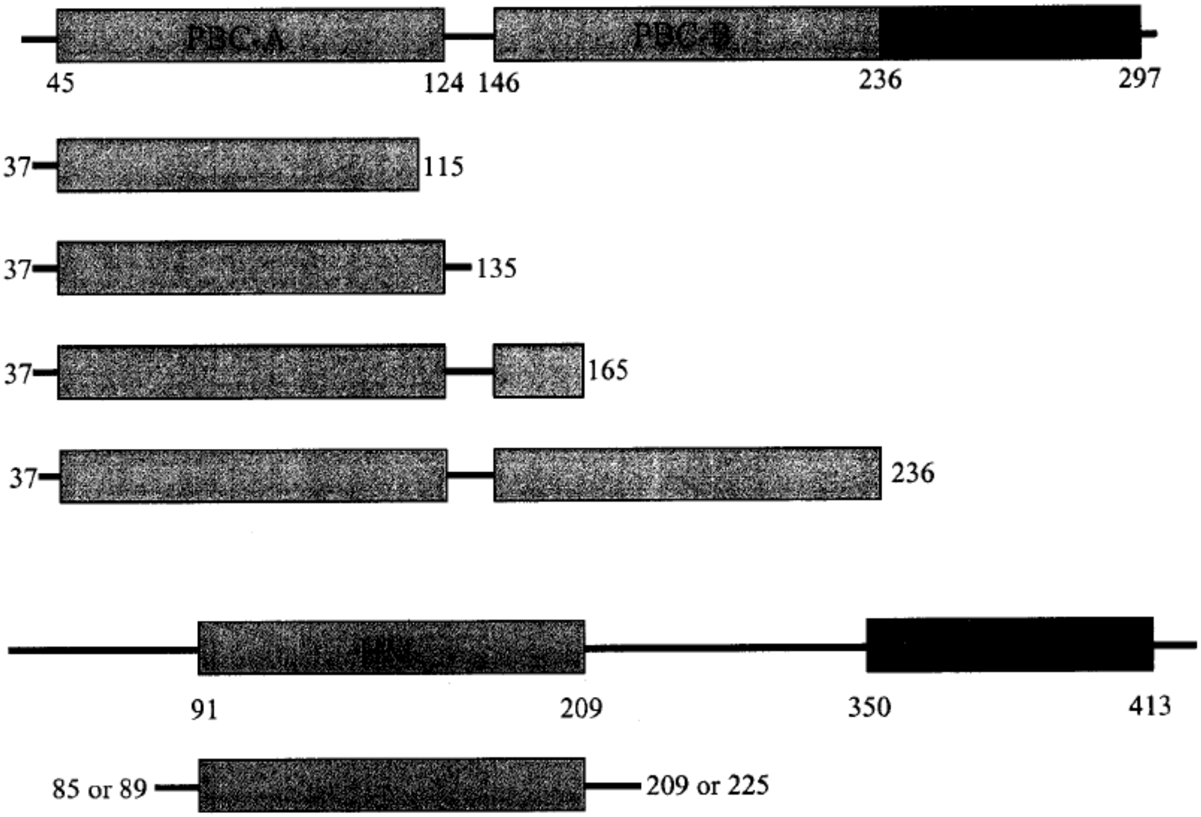


Figure 5.2: **Partial proteolysis of Homothorax and Extradenticle**
Results of the partial proteolysis of the Hth/Exd complex carried out by Keri Siggers in the context of her PhD thesis [22]. Exd is shown at the top, Hth at the bottom. Black boxes represent homeodomains. Reprinted from [22] with permission from Keri Siggers, PhD.

5.2 Formation binary, ternary and quarternary complexes

5.2.1 Optimization of constructs, expression and purification protocols: ScrHD, ExdHD and HM/Exd

Guided by Keri Siggers' proteolysis data, secondary structure prediction programs (PSIPRED [119]) and hydrogen-deuterium exchange (HDX) mass spectrometry conducted by our collaborator Dr. Gaetano Montelione, we screened a number of constructs for both HM and

Table 5.1: **Screened constructs for HM and Exd**

HM construct	Exd construct	Soluble?
FL-HM (1-247)	FL-Exd (1-376)	**Yes**
HM 15-247	Exd 1-315	**Yes**
HM 100-224	Exd 37-311	No
HM 79-247	**Exd 37-311**	**Yes**
HM 79-198	Exd 37-311	No
HM 79-203	Exd 37-311	No
HM 79-207	Exd 37-311	No
HM 79-212	Exd 37-311	No
HM 79-222	Exd 37-311	No
HM 79-225	Exd 37-311	No

Exd (table 5.1). The most important result from these screens was that HM and Exd need to be coexpressed from the same plasmid to avoid precipitation and aggregation. The shortest constructs for both proteins that led to a soluble 1:1 expression was obtained by a pET-Duet vector that included residues 79-247 for HM and 37-311 for Exd, driven by a single T7 promoter. Interestingly,residues 198-247 seem to be disordered according to secondary structure prediction, proteolysis and HDX analyses but turn out to be necessary for solubility of the HM/Exd complex when expressed from the pET-Duet vector.

The complex was then expressed and purified by simply tagging the HM N-terminus with a His$_6$-tag, followed by ion-exchange and/or size exclusion chromatography. Binding between the two proteins is very strong such that they co-elute as a complex. This complex of HM 79-247 and Exd 37-311 will simply be called HM-Exd from here on.

For further details about this optimization process please refer to Chapter 2 (Materials and Methods).

5.2.2 Magnesium is needed for proper complex formation

Mixing ScrHD with ExdHD or HM-Exd in the absence of DNA leads to instant precipitation. The same happens when mixing DNA with any of the protein components. This agrees with the issues outlined in Nichole O'Connell's thesis [23]. The pI's of the Scr and Exd homeodomains are 11 and 10, respectively, due to a large number of basic amino acids, needed for non-specific interaction with the DNA backbone before formation of the more specfic interactions with the DNA basepairs. This electrostatic interaction is very probably the reason for the precipitation seen at the high concentrations of protein and DNA needed for biophysical studies.

Previous screens of mixing conditions performed by Dr. O'Connell did not include divalent cations like magnesium, that are known to screen the phosphate backbone of nucleic acids. After again screening different buffer conditions, we found that magnesium seems to be the most important variable when it comes to avoiding precipitation upon mixing DNA with any of the protein components. Interestingly, preliminary tests indicate that once Scr and DNA have been mixed in the presence of a high concentration of magnesium, one can reduce the concentration of magnesium again without causing precipitation. This could mean that high magnesium concentrations are only needed to screen the phosphate backbone during the initial "shock" of mixing the highly concentrated protein and DNA fractions together, but can be reduced once most of the complex formation has happened (displacing much of the magnesium from the DNA) and local concentrations of the free constituents is smaller. Since many commercially available $MgCl_2$ salts have paramagnetic contaminants like manganese that increase NMR lineshapes, an $MgCl_2$ salt of >99.995% purity was used when preparing the complex for NMR spectroscopy.

Further addition of ExdHD equilibrated in $MgCl_2$ resulted in good solubility and no further precipitaition. However adding HM-Exd resulted in further precipitation over time.

This could mean that more magnesium is needed in the case of the larger complex or that other mechanisms, such as proteolysis or conformational changes leading to aggregation, are at play.

5.2.3 Gel shifts, SELEX-Seq and target DNA sequences

Slattery et al. [120] developed an experimental/computational platform called SELEX-Seq, which conveniently allows identification of a large number of target sequences from a DNA library for each HOX protein together with relative affinities. The article groups the resulting sequences in 10 main classes (for convenience named after the color scheme used in the research article, e.g. "red" and "blue" sequences) that correspond to different core sequences of the HOX-Exd binding motif and different affinities (and thus specificities) to the eight drosophila HOX proteins. The red group for example corresponds somewhat to a consensus group with high affinities among many HOX proteins, especially posterior ones (low specificity). Whereas the "blue" group has high affinities for more anterior HOX proteins, namely Scr and Dfd (high specificity).

The "blue" DNA core 12mer described herein ("blue12": ATGATTAATTGC) corresponds the top Scr target sequence from the SELEX-Seq 12mer dataset, while showing less affinity to more posterior Hox proteins (relative affinities of 1.0 and 0.27 for Scr and Ubx-IVa, respectively). The "red" core 12mer described herein ("red12": ATGATTTATGAC) on the other hand has relatively high relative affinities for multiple Hox proteins (0.93 and 0.67 for UbxIVa and Scr, respectively).

Complex formation was accessed in three different ways. Gel shifts performed by Namiko Abe and Katherine Lelli in the Mann lab show cooperative binding of Scr and HM/Exd constructs described herein of both the red and blue core sequence when flanked by a GC on both sides (red16 and blue16). The blue12 core sequence did not show any binding. This shows that both Scr and HM-Exd constructs as well as the blue16 sequence are viable for NMR binding studies. In chapter 6 we will see that 14mer DNA oligomers are sufficient for cooperative binding of Hox and Exd homeodomains.

Additionaly, ExdHD as well as HM/Exd were mixed with Scr and DNA under the conditions described before at approximately equimolar amounts (with about 20% excess DNA) and run over a size exclusion column (Superdex S200). If a stable complex is formed, then the proteins should elute together in a single peak of the approximate molecular weight of the complex instead of several peaks corresponding to its constituents. This was indeed the case for both the smaller (blue16-ScrHD-ExdHD) and larger complex (blue16-ScrHD-HM/Exd), which eluted in single peaks around 30 and 70 kD, respectively, as expected, with a gel confirming the presence of all protein components at approximately equimolar amounts.

Lastly, we assessed complex formation by NMR spectroscopy as described in the following sections.

5.3 One- and two-dimensional NMR spectroscopy of the DNA moiety

One of the first NMR spectra we collected was a simple 1D "Jump Return" experiment looking at the imino protons of the DNA (blue16 unless otherwise specified). The spectrum shows 14 peaks - 1 for each base pair except the two terminal base pairs (invisible due to fast exchange with the water protons of the solvent). The more downfield peaks, corresponding to the AT region, are more crowded leading to more spectral overlap (see Figure 5.4). This is unfortunate as this is the core binding site for the two homeodomains (the blue16 sequence is GCATGATTAATTGCGC, with the Exd-Scr binding motif highlighted). Since complex formation will broaden the lines due to spin relaxation, this will make the spectrum even harder to interpret (see below).

Next, a 2D NOESY spectrum was collected on our in-house 500 MHz spectrometer, to try and assign the 14 peaks, but strong overlap of the peaks on the "A/T" region made assignments within that region difficult. After collecting additional 2D NOESY spectra on a 900MHz instrument at the New York Structural Biology Center (NYSBC) we were able to assign all 14 imino peaks to the respective base pairs in the DNA sequence.

5.4 Measurement of transverse relaxation times to assess complex sizes

To further assess complex formation, we performed a series of 1D experiments of blue16 in the presence of the different protein components to determine linewidths and transverse relaxation times, T_2. This allowed us to estimate tumbling times of the different complexes and thus apparent complex sizes. Salt titrations from 90 to about 500 mM NaCl were then performed on each complex to asses binding strength. Increasing the ionic strength of a solution will weaken electrostatic interactions according to Debye-Hückel theory and thus lead to dissociation of any macromolecular complexes that are held together mainly by electrostatic interactions. The studied samples were blue16 alone, blue16-ScrHD, blue16-ExdHD, blue16-ScrHD-ExdHD ("Scr-Exd-DNA" or "SED"), blue16-HM/Exd, and blue16-HM/Exd-ScrHD ("Scr-HM/Exd-DNA" or "SHED").

Across the board, adding protein to the DNA increased the linewidths of the imino peaks suggesting interaction with the proteins (Figure 5.5). An improved pulse sequence using a so-called Hahn Echo to refocus inhomogeneous evolution of spins due to chemical shifts and field inhomogeneities allowed us to measure transverse relaxation (T_2) more directly, while also getting much better water suppression and thus better base lines and line shapes which were easier to fit (Figure 5.6, pulse sequence adapted from [30]). Salt titrations with this new Jump Return sequence with Hahn Echo reveal that while Scr and blue16 almost completely dissociate at 500mM NaCl (T_2's having plateaued at values similar to those of DNA alone), Exd seems to bind more strongly and the T_2's do not quite reach those of free DNA at 500 mM salt.

The ratio of T_2's at 500 and 90 mM NaCl of Exd-blue16 is very close to the ratio of molecular weights of 1.9 while the ratio for Scr-blue16 is around 3. Considering that salt titration for Scr-blue16 seems to have reached a plateau while the one for Exd-blue16 has not, it seems reasonable to assume that final ratios for Exd-blue16 will also be closer to 3. Indeed, assuming the same final T_2 values (which should be similar assuming we are seeing only unbound DNA at the end of the titration for both) the ratio is around 3 as well.

T_2's at 90 mM salt (when the complex is presumably stably formed) are around 5-7 ms for Scr-blue16 and 6-8ms for Exd-blue16, which indeed is slightly smaller (see Table 5.2). T_2's of the ternary complex SED are around 3-4 ms, those of free DNA around 12-18 ms depending on the position of the base pair in the sequence. The ratio of T_2's at 500 and 90 mM NaCl for SED is around 3.5 but since here too values did not reach a plateau, one has to assume that the final value will be closer to 5.

Some of the discrepancy between T_2 ratios and molecular weight ratios might be due to the differences in salt concentrations affecting relaxation times, for example through changes in water structure or changes in structure and dynamics of the free DNA at high salt. Indeed some reports in the literature suggest a decrease in the radius of gyration of DNA at high salt concentrations [121]. When performing the salt titration experiment on free DNA, we indeed see an increase in relaxation times of free DNA at high salt. Using the transverse relaxation times of free DNA at 90 mM salt for the calculation of the ratios, yields values closer to the expected molecular weight ratios, albeit still slightly high (see

Table 5.2).

When plotting the transverse relaxation rates R_2 ($1/T_2$) against the expected molecular weights, a linear relationship is seen (Figure 5.7). While the linear relationship was expected for globular molecules (Stoke's law), we were surprised to see a negative zero-crossing. One plausible explanation is an overestimation of the slope due to effects of protein binding on the shape of the complex and thus the tumbling behavior. Rotational diffusion rates along different axes of a complex are different depending on the shape of the complex. Binding of a homeodomain on the DNA will have different effects on the different rotational tumbling rates. In particular, because the imino bond vector is approximately perpendicular to the long axis of the DNA, it seems plausible that the relaxation rates increase more than expected for increasing the molecular weight isotropically, that is without changing the shape of the molecule. To answer this question, theoretical models for tumbling behavior in solution could be established based on the known three dimensional structures of the components of the complex.

An additional explanation could be an additional rigidification of the imino protons within the DNA duplex upon binding of the homeodomains. In the case of the ternary "SED" complex, whose T_2 ratio exceeds its MW ratio even more than for the other complexes, it seems plausible to think that some parts of the proteins that are disordered for the binary complex, become ordered in the ternary complex (the Scr NTA/linker region is a good candidate, as it binds Exd). To test these hypotheses, additional T_1 and T_2 relaxation experiments can be performed on the protein moieties, in particular with ^{15}N as probe nucleus, instead of the chemically more labile imino protons.

Table 5.2: **Transverse relaxation times for blue16 in the absence and the presence of Scr and Exd homeodomains**

Peak	blue16	blue16 +ScrHD	Ratio	blue16 +ExdHD	Ratio	“SED”	Ratio
G15	11.66	5.34	2.19	6.59	1.77	N/A	N/A
G14	18.01	12.73	1.41	7.18	2.51	N/A	N/A
G2	11.54	6.43	1.79	6.17	1.87	3.49	4.93
G13	17.20	5.83	2.95	7.67	2.24	3.70	4.65
G5	17.19	6.28	2.74	6.42	2.68	3.59	4.01
Average	18.74	5.97 (w/o G14)	2.42	6.81	2.21	3.59	4.53
MW(kDa)	9.8	21.2	-	19.7	-	31.1	
$\frac{MW_{complex}}{MW_{blue16}}$	1.00	-	2.17	-	1.92	-	3.10

Transverse relaxation times for blue16 alone or in the presence of Scr, Exd or both homeodomains. G14 is a strong outlier in the case of blue16-ScrHD and was omitted from averaging. Ratios of relaxation times to the relaxation time of free DNA exceed the molecular weight ratios by a factor of 1.1-1.4.

5.5 Fast to slow exchange transition suggest cooperative binding

The fact that the SED complex is harder to dissociate with increasing salt is a sign of cooperative binding. Another sign of cooperative binding is the fact that the complex seems to be in slow exchange with free DNA as opposed to the binary complexes, which seem to be in fast exchange.

In the case of blue16-ScrHD and blue16-ExdHD increasing the salt concentration gradually moves the peaks downfield while simultaneously gradually getting sharper speaking in favor of the bound and unbound forms being in fast exchange. In the ternary SED complex, however, the peaks initially do not change when increasing the salt concentration. At around 300mM salt additional sharper peaks appear further downfield. At even higher salt concentration the broader upfield peaks disappear and only the sharper downfield peaks remain.

This behavior speaks in favor of the ternary complex being in slow exchange. Since for both homeodomains alone we seem to see fast exchange it seems like binding is much stronger in the presence of both homeodomains as compared to either one being present alone, indicating cooperative binding of the homeodomains. This is likely due in part to the previously described binding of the YPWM motif of Scr to the TALE motif of Exd but could also be due to more indirect cooperative effects such as binding of one homeodomain affecting the shape of the DNA in such a way that binding of the second one becomes more likely. Evidence of DNA shape changes upon binding is presented in the following chapter (chapter 6). This idea can be tested more thoroughly by repeating these experiments with an NTAless version of Scr or mutants of Scr or Exd incapable of interacting through the YPWM-TALE interaction (for example a mutation in the YPWM or TALE motifs), to see how much, if any, cooperativity is lost.

5.6 HM-PBC domains might be tumbling independently from Exd homeodomain

We also carried out the salt titration experiments for blue16 with HM/Exd. While the positions of the peaks were different from the peak positions of blue16 with only the homeodomain of Exd, both the linewidths as well as the transverse relaxation times were very similar, possibly indicating that the PBC and HM domains tumble relatively independently from the DNA and homeodomain (Table 5.3). It remains to be seen if PBC and HM domains expressed independently can still interact with the homeodomain of Exd, but if this is the case it would provide us with some powerful tools to study the effect of PBC-HM on the interaction of Exd with Scr and the DNA by titrating PBC/HM into a prebound SED complex. In addition this possibly means that PBC/HM on its own can adopt a near native confirmation, allowing structural studies by NMR or X-ray crystallography separately from its homeodomain and DNA. Both PBC and HM domains structures are unknown and would be the first of their respective homology groups to be characterized structurally.

Finally, we tried to add ScrHD to the prebound complex of blue16 and HM/Exd. While no instant precipitation was visible, as it appears when mixing components in the absence of magnesium, precipitation appeared after several minutes. This could simply mean that the larger complex needs more magnesium for stable interaction or that proteolytic and/or conformational changes on the time scale of minutes are leading to precipitation. Very surprisingly the linewidths and transverse relaxation times of the imino peaks again are similar to those of blue16 with only one homeodomain present, while the appearance and peak positions is different from any of the other spectra collected. Possibly we are seeing some type of mixture of the different spectra, but it seems unlikely that we are seeing a quarternary complex because the T_2's are much larger than expected for two homeodomains bound to the DNA (5-8 ms as opposed to 3-4 ms for SED), suggesting that any quarternary

Table 5.3: **Transverse relaxation times for blue16 in the presence of HM/Exd**

Peak	blue16	blue16 +ExdHD	Ratio	blue16 +HM/Exd	Ratio
G15	19.36	6.59	1.77	6.13	1.90
G14	18.50	7.18	2.51	6.54	2.75
G2	18.71	6.17	1.87	5.00	2.31
G13	19.23	7.67	2.24	6.54	2.63
G5	17.90	6.42	2.68	5.19	3.31
Average	18.74	6.81	2.21	5.88	2.58
MW(kDa)	9.8	19.7	-	59.9	-
$\frac{MW_{complex}}{MW_{blue16}}$	1.00	-	1.92	-	5.14 (1.92 HD only)

The ratio of transverse relaxation times of the blue16-HM/Exd complex to free DNA is very similar to that of blue16-ExdHD, suggesting that the HM and PBC domains (Exd without homeodomain) tumble relatively freely in solution.

complexes have fallen out of solution.

Another explanation is that the NTA of Scr, which we know to be prone to cleavage, is not present at the time of the experiment and we thus saw a mix of blue16-ScrHD and blue16-HM/Exd spectra, which both have similar line widths and relaxation times as the ones we observe.

5.7 Identification of residues involved in binding and conformational changes from two-dimensional NMR spectroscopy

5.7.1 ^{1}H/^{15}N-backbone-amide experiments of Scr

We collected preliminary 2D backbone amide spectra of ^{2}H/^{15}N-labeled Scr in the absence or the presence of DNA. The overall look of the spectrum was very similar to similar spectra collected previously by Nichole O'Connell [23], but many individual peak positions were different, presumably due to buffer and temperature differences. Spectra collected of Scr N321D ("ScrND"), a mutant mimicking the deamidated version of Scr as described in Nichole O'Connell's thesis, are virtually indistinguisgable from the spectra for Scr. Upon addition of DNA a large number of peaks in the spectrum shifted suggesting interaction with the DNA (Figure 5.9).

Addition of Exd homeodomain to Scr-blue16 did not seem to affect the spectrum much in preliminary experiments, whereas addition of HM-Exd had a strong effect on the spectrum that might be due to protein aggregation as suggested by the strong precipitation seen in that sample.

These experiments were conducted in the absence of $MgCl_2$, which is needed for optimal interaction and solubility as described earlier. Additionally, experiments were conducted with samples of different ages and cast some doubt on the different degrees of degradation of the involved proteins, in particular for Scr, whose N-terminal linker region is important for complex formation and prone to degradation. For proper interpretation these experiments need to be repeated with fresh and clean protein samples in the presence of $MgCl_2$, using 1D imino spectra to assess binding in parallel of the 2D amide backbone spectra.

Nonetheless, several peaks corresponding to amino acids known to be important for the interaction can be shown to undergo similar shifts in all of the spectra upon addition of DNA, which in combination with the imino-proton spectra and the gel shift assays give rise to hope that specific binding is being seen.

5.7.2 ^{13}C-methyl TROSY of Scr

Next, we expressed and purified an ILV (Isoleucine, Leucine, Valine) ^{13}C-methyl labeled Scr homeodomain. There is a number of advantages of using methyls as probes: the degeneracy of their three protons effectively increases the concentration of each group for NMR purposes and the rotation of the methyl group (three site hop) and their usual position at the ends of side chains makes them relax more slowly further increasing the signal-to-noise and resolution. Additionally, the spectrum will be less crowded than for a fully labeled protein and thus easier to analyze.

We were able to collect very nice preliminary methyl spectra for the Scr homeodomain at 600 MHz (Figure 5.10). Repeat experiments were carried out at NYSBC at 800 MHz and resulted spectra with lower signal-to-noise but nearly identical peak positions (Figure 5.11). The same experiment was then repeated in the presence of DNA and then of both DNA and ExdHD. A few peaks are changing upon addition of DNA and then again upon addition of ExdHD.

The spectrum can be subdivided in three regions that based on the chemical shift predictor ShiftX2 [122] correspond to either peaks for Leucines, Valines or Isoleucines (Figure 5.10). Upon addition of DNA we can see a number of changes in the Leucine region, indicating some rearrangements in the hydrophobic core of the protein. There are also some changes in the Isoleucine region, in particular for two peaks. Two Isoleucines lie in the so-called recognition helix 3 that binds to the major groove of the DNA, one of them forming direct hydrogen bonds to the DNA bases. Upon addition of Exd, changes in the hydrophobic core (Leucine region) are minor, but new peaks appear both in what we believe to be the Valine region and the Isoleucine region. There is an Isoleucine right next to the YPWM motif which is known to bind to Exd's TALE motif. It is very possible that the new peak appearing at the upper end of the spectrum corresponds to that Isoleucine. Additionally, new peaks appear in the Valine region, which would suggest some conformational change in the linker of Scr, where both Valines lie. This seems plausible because the linker connects the homeodomain of Scr to its YPWM motif, which interacts with Exd. The Leucine region in particular shows some peak doubling, which we believe to be due to slow cleavage of the NTA, leading to the mixture of spectra for Scr-blue16 and free Scr (due to part of the DNA in solution now being occupied by Exd alone).

For proper interpretation, assignments of the ILV methyls need to be done using a fully ILV-sidechain and backbone labeled Scr sample. Also these experiments should be repeated at higher concentrations to reduce the time of the experiment and with larger amounts of protease inhibitor present, both to reduce the influence of proteolysis on the spectra.

5.8 MD simulations of Scr and Scr mutants with fkh and fkhCON DNA

We conducted molecular dynamics (MD) simulations on Scr and several Scr mutants mimicking its paralog Ubx with and without the specific target sequences fkh (which binds Scr only) and fkhCON (which binds Scr and Ubx), whose crystal structures were solved in complex with Scr in 2007 [49]. These simulations show some interesting features that could possibly explain some of the affinity differences between Scr and Ubx, and thus specificity.

As shown by Joshi et al. [49], the insertion of Arginine 3 into the minor groove is essential for the specificity of Scr to its target sequence fkh. According to the crystal structures solved in the publication, Arg3 of Scr inserts into a local width minimum of the minor groove, thus being used to "pick out" DNA targets with particularly narrow minor grooves at this spot. Interestingly, the study also points out that other Hox proteins that cannot bind to fkh, still have an Arginine at position 3, but differ at the closeby positions 4 and 6. While Scr has a Glutamine and a Threonine at positions 4 and 6, Ubx - which cannot bind fkh - has a Glycine and a Glutamine at these positions, yielding the motif RQRT for Scr and the motif RGRQ for Ubx.

We simulated the homeodomain of wild-type Scr bound to fkh and its consensus variant fkhCON, as well as three mutants of Scr mimicking Ubx at positions 4 and 6. Namely the two point mutants Scr Q4G and Scr T6Q and the double mutant Scr Q4G/T6Q. We then studied Arg3's occupancy of the canonical "inserted" state, by measuring the distance of its Guanidinium group to Adenine 13 inside of the minor groove (Figure 5.12).

In the trajectories of wtScr with its specific target fkh, Arg3 seems to spend more time in the canonical, low entropy state, with its side chain extended and inserted into the minor groove than in the simulations with fkhCON. In the case of the wtScr trajectory with fkh, this measure populates three states: 1) inserted into the minor groove (majority of the time) 2) the Guanidinium group bending back on itself to interact with the phosphate backbone of the DNA and 3) the entire residue coming out of the minor groove and the N-terminal amine interacting with the phosphate backbone of the DNA. The latter state is very probably falsely inflated, because the simulated construct starts at residue 3, which is not the native N-terminus. In the case of wtScr on fkhCON a fourth state is observed, which we name 2a, with the Guanidinium group interacting with Arg3's own carbonyl group, which is populated about 50% of the time. Figure 5.13 is a graphical representation of Arg3's the population of these four states in the different trajectories.

The mutant ScrQ4G seems less confined to state 1 and spends more time in state 3, presumably because the Glycine at position 4 allows for more flexibility of the N-terminus. A rerun of the same simulation however, shows Arg3 confined to state 1 for the entirety of the trajectory. This presented us with somewhat of a puzzle, but when analyzing Gly4's backbone dihedral angles, we noticed that Φ of Gly4 undergoes an 180° rotation early in the trajectory, that leads to non-native hydrogen bonds of Gly4 with the backbone of the DNA, that we believe artificially increases Arg3's population of state 1.

Something interesting was observed when analyzing the simulation of the second mutant Scr T6Q. In the case of wtScr, Threonine 6 forms a hydrogen bond with the phosphate backbone of the DNA, which is weakened in Ubx where the Threonine is replaced by a Glutamine. Threonine 6 is stably hydrogen bonded to the backbone phosphate for the majority of the trajectory, whereas Glutamine 6 is rather mobile and constantly makes and breaks its hydrogen bond with the DNA backbone. When examining a number of crystal structures of Scr and Ubx bound to DNA it seems that Thr6 is indeed in hydrogen bonding distance to the phosphate backbone for several of them, while Gln6, which takes its place in Ubx, seems to usually be disordered and not show any clear electron density, as suggested by the simulations. When analyzing the trajectory of wtScr on fkh, there seems to be a strong correlation between the breakage of the hydrogen bond of Thr6 to the DNA backbone and Arg3 leaving its “inserted” state 1 (see Figure 5.14). The same is true for the trajectory of Scr Q4G. In the case of Scr T6Q, however, the hydrogen bond of Q6 with the DNA backbone is frequently broken and seems uncorrelated with Arg3 insertion into the minor groove.

An additional result of the examination of these trajectories is that all simulations with fkhCON as well as both simulations containing the T6Q mutation (Scr T6Q, Scr Q4G T6Q), show the additional “2a” state of Arg3 being populated (see Figure 5.13). This new state corresponds to the Guanidinium group of Arg3 bending back to interact with its own backbone carbonyl. Our interpretation of these last two observations is that the wider groove of fkhCON allows for Arg3 to adopt a higher entropy state and explore conformations that would be strained in the case of Scr’s native target fkh. Mutating Threonine 6 to Glutamine seems to take away some of the constraints that the narrow minor groove of fkh imposes on Arg3. Indeed breaking of the hydrogen bond of Thr6 with the phosphate backbone seems to show a strong negative correlation with the insertion of Arg3 into the minor groove (see above). In the T6Q mutant this correlation seems to be abolished. Thr6 has previously been identified as one of the residues conferring specificity to Scr [118]. The T6Q mutation thus makes Scr more Ubx/Antp-like and thus should decrease its affinity to fkh.

5.9 Discussion

The work discussed in this chapter builds upon the work conducted by Keri Siggers and Nichole O’Connell as part of their respective PhD theses on the Hox protein Scr and its cofactors Exd and Hth (or its HM domain). Great progress has been made here in identifying constructs and expression and purification schemes for HM/Exd in particular, giving future researchers a solid basis to study these proteins. In addition, identifying the right conditions for complex formation as well as a quick NMR experiment to assess complex formation, will be of great use for future NMR studies of Scr, Exd and DNA, in particular because no isotope labeling is needed to collect imino spectra. Finally, assignments of the imino peaks has been performed for blue16, making it a good probe for future experiments on the system.

Measurements of transverse relaxation times T_2 yielded very promising results for the formation of specific complexes of blue16 with the homeodomains of Scr and Exd. In addition, salt titration experiments strongly suggest cooperative binding of the two homeodomains in the case of the ternary complex, because of strongly reduced dissociation at high salt concentrations. This is further supported by the fact that the ternary complex, as opposed to either of the binary complexes, shows to be in slow exchange with free DNA. The salt titrations further indicate that much of the binding is electrostatically driven.

Ratios of T_2 values of the respective complexes to the values of free DNA slightly exceed the ratios of molecular weights of the complexes to free DNA. Several hypotheses could explain this behavior, including but not limited to non-linear increase of tumbling times with molecular weight, presence of additional relaxation mechanisms for imino protons, especially for unbound DNA, as well as disorder-to-order transitions in the protein components, especially for the ternary complex. It seems more than plausible to assume that the linker region of Scr gets somewhat more ordered upon interaction with Exd. All these issues can be addressed partly by performing additional relaxation experiments (T_1 and T_2) on the protein moieties instead of the DNA moiety of the complex, using ^{15}N as probes, instead of the rather labile imino protons.

While much of this indicates cooperative binding, it does not prove that cooperativity is solely due to the YPWM-TALE interaction. To study this, one would need to repeat these experiments with Scr mutants that do not contain the linker, and/or Scr or Exd mutants lacking the ability to interact through these motifs. Other mechanisms might contribute to cooperativity, such as changes in DNA shape upon binding of the first protein (see chapter 6 for a discussion on this).

One very interesting result of this chapter is that the apparent tumbling time of HM/Exd bound to DNA seems to be very similar to that of only the homeodomain of Exd bound to the DNA. This possibly indicates that the PBC-HM domains of HM/Exd tumble relatively independently from the homeodomain and thus the DNA in solution. This is important for several reasons:

1) It reduces some of the burden of studying a high molecular weight complex by NMR, because the complex relaxes like a lower molecular weight complex.

2) It probably means that HM/PBC are relatively stable without the homeodomain of Exd, possibly allowing them to be expressed and studied independently. Neither the HM nor the PBC domain have been characterized structurally, nor has any homologous domain.

3) If PBC-HM can be indeed expressed and purified independently, they might still interact with a preformed DNA-ExdHD complex. This would open up the possibility to study the effect of HM/PBC on the canonical DNA-homeodomain complex, by titrating PBC/HM to a preformed DNA-HD complex.

Lastly, a few preliminary 2D spectra have been collected for Scr, both backbone amide labeled and side chain methyl labeled. While more work clearly needs to be done here, these preliminary experiments support some of the conclusions from the other experiments. Namely, adding DNA to both the backbone and the methyl labeled Scr samples causes peak shifts, further supporting complex formation. While the spectra for the backbone experiments for the ternary complex are inconclusive, the methyl spectra clearly indicate further peak shifts, arguing in favor of a direct interaction of Exd with Scr. Especially new peaks appearing that we believe to belong to the valines (which both lie in the linker region of Scr, which is known to interact with Exd) and a peak appearing in the isoleucine region (because an isoleucine lies adjacent to the YPWM motif of Scr) give rise to the hope of indeed seeing cooperative interactions between Scr and Exd.

The results of the MD simulations of Scr and its mutants with fkh and fkhCON are very promising and provide us with some interesting hypotheses about the differences in affinities of Scr and Ubx, and the importance of residues 4 and 6 for specificity. One very interesting result that came out of these simulations was the fact that Threonine 6 and a DNA backbone phosphate form a strong hydrogen bond, which shows a strong correlation with insertion of Arg3 into the minor groove. Mutating this residue to a Glutamine not only weakens this hydrogen bond and all but nullifies the correlation, but also seems to lead to Arginine 3 being able to occupy a new conformation, interacting with its own backbone. This conformation can only be seen in simulations of the Scr mutants containing this T6Q mutation with fkh and in all simulations with fkhCON. This suggests that the hydrogen bond of Threonine 6 is needed to keep Arginine 3 in place, and that the energy from this hydrogen bond makes up for the entropy loss of inserting Arg3 into the narrow minor groove in the case of fkh, but makes little or no difference in the case of fkhCON where Arg3 is not as confined to begin with, due to a wider minor groove. This ties in well with the work of Abe et al. [118], which found Threonine 6 to be of particular importance for the shape readout functionality of Arginine 3. These results are further underlined by the structure comparisons in chapter 6.

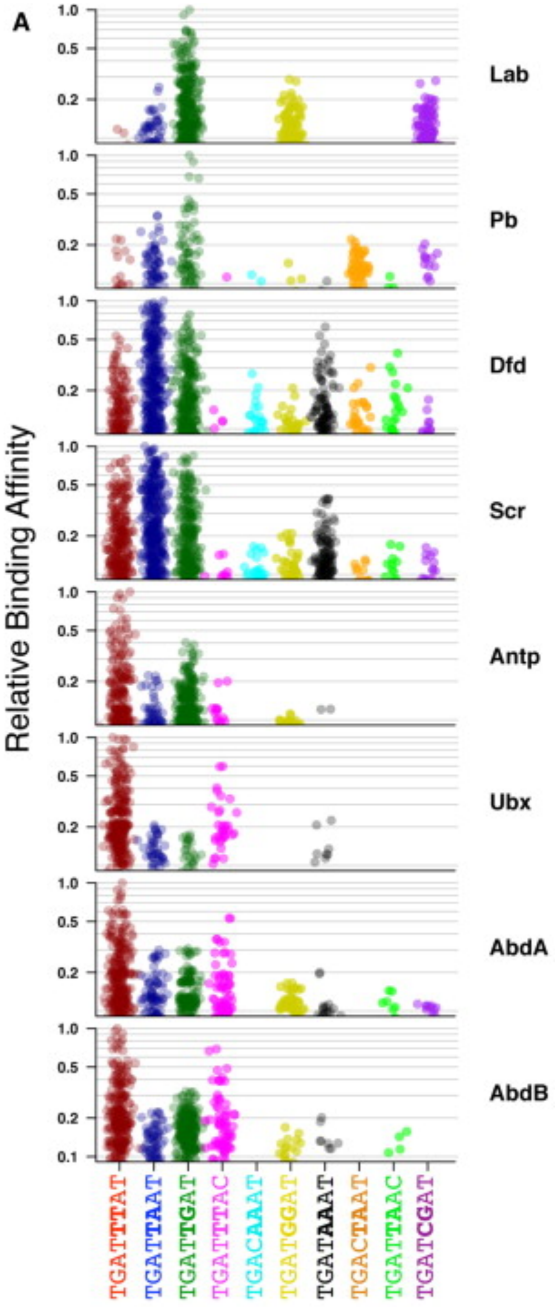


Figure 5.3: **Ten core binding motifs revealed by SELEX-Seq**
The eight *Drosophila* Hox proteins and their relative affinities to the 10 core binding motifs identified by SELEX-Seq. Figure reprinted from Cell, 147, Slattery M, Riley T, Liu P, Abe N, Gomez-Alcala P, et al. (2011) Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. 1270 - 1282. Copyright 2011, with permission from Elsevier. [120]

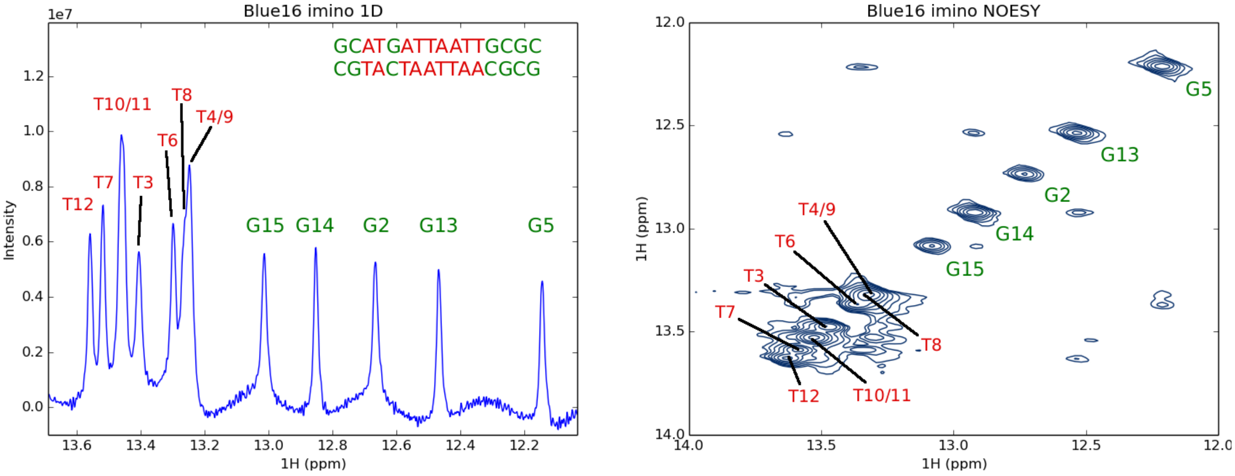


Figure 5.4: **1D and 2D spectra of imino protons for blue16**
The left panel shows a 1D spectrum of the imino region for a jump return experiment collected on blue16 on our in house 500 MHz spectrometer. The right panel shows the corresponding region in a 2D NOESY experiment for blue16, collected on the same spectrometer. Assignments were done with the help of spectra at different fields, collected in-house and at the New York Structural Biology Center. As can be seen in the NOESY experiment a number of peaks overlap.

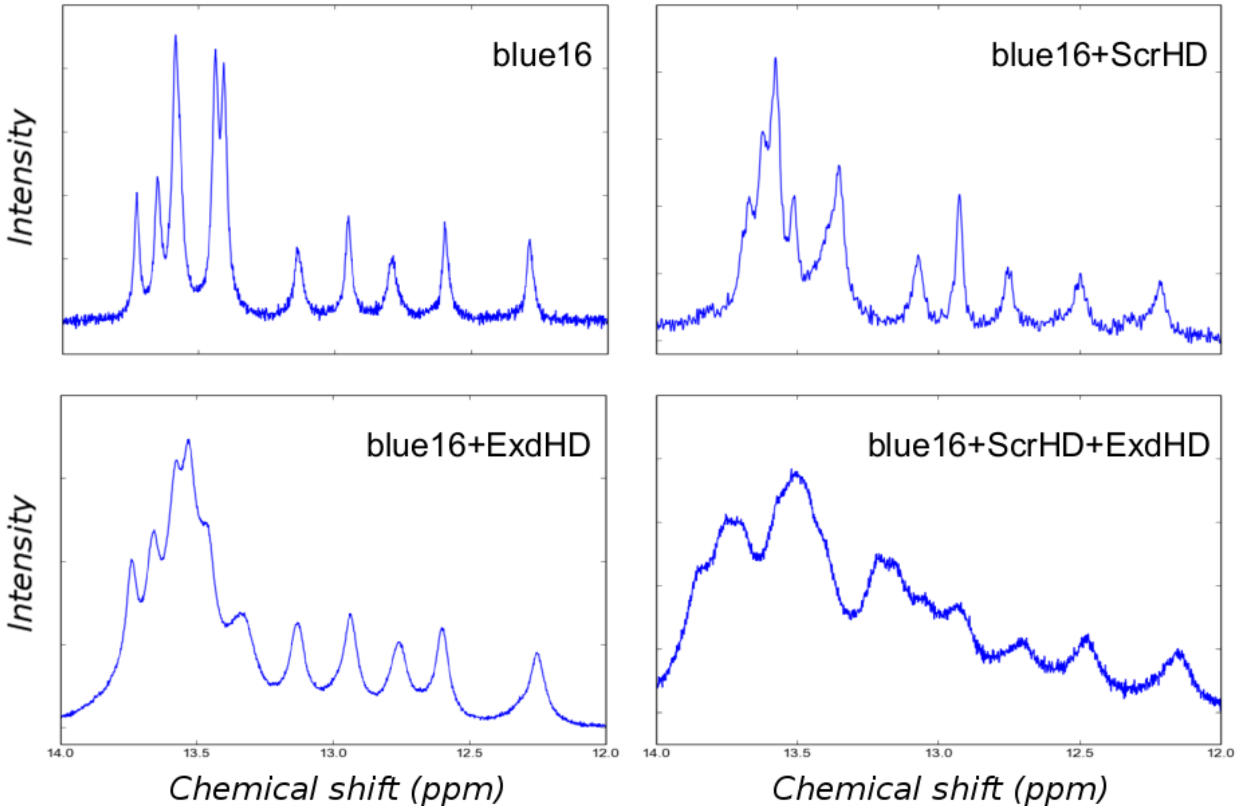


Figure 5.5: **1D spectra of the imino region for various complexes**
Imino spectra for four different samples: blue16, blue16+ScrHD, blue16+ExdHD and blue16+ScrHD+ExdHD ("SED" complex). Lineshapes increase with molecular weight as expected, speaking in favor of complex formation.

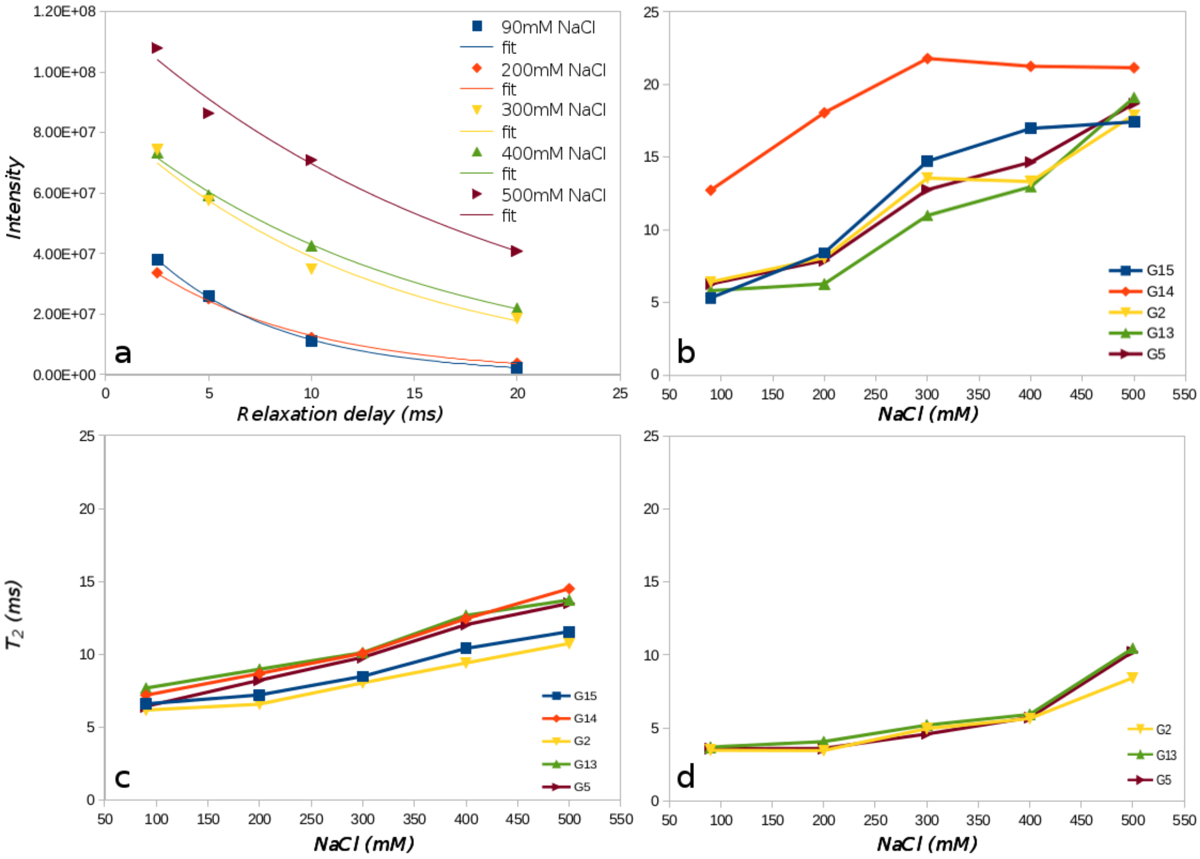


Figure 5.6: **Salt titrations and transverse relaxation times**
Salt titrations and calculation of transverse relaxation times for three complexes. a) Intensities of Guanine peaks as a function of relaxation delay for blue16-ScrHD. Fitting of an exponential function yields the transverse relaxation times T_2. The represented curves correspond to Guanine 5. b) Transverse relaxation times T_2 as a function of salt concentration for blue16-ScrHD. Relaxation times clearly increase with concentration of NaCl speaking in favor of a dissociation of the complex. Final relaxation times at 500 mM NaCl are close to those of free DNA. c) T_2 as a function of salt concentration for blue16-ExdHD. Relaxation times increase with salt but do not reach the same levels as free DNA or blue16-Scr at 500 mM, suggesting incomplete dissociation. d) T_2 as a function of salt concentration for the ternary SED complex (blue16-ScrHD-ExdHD). Dissociation is even slower and less complete than for blue16-ExdHD speaking in favor of cooperative binding. The peaks for G15 and G14 were omitted because they could not be fit due to overlap with peaks in the thymine region.

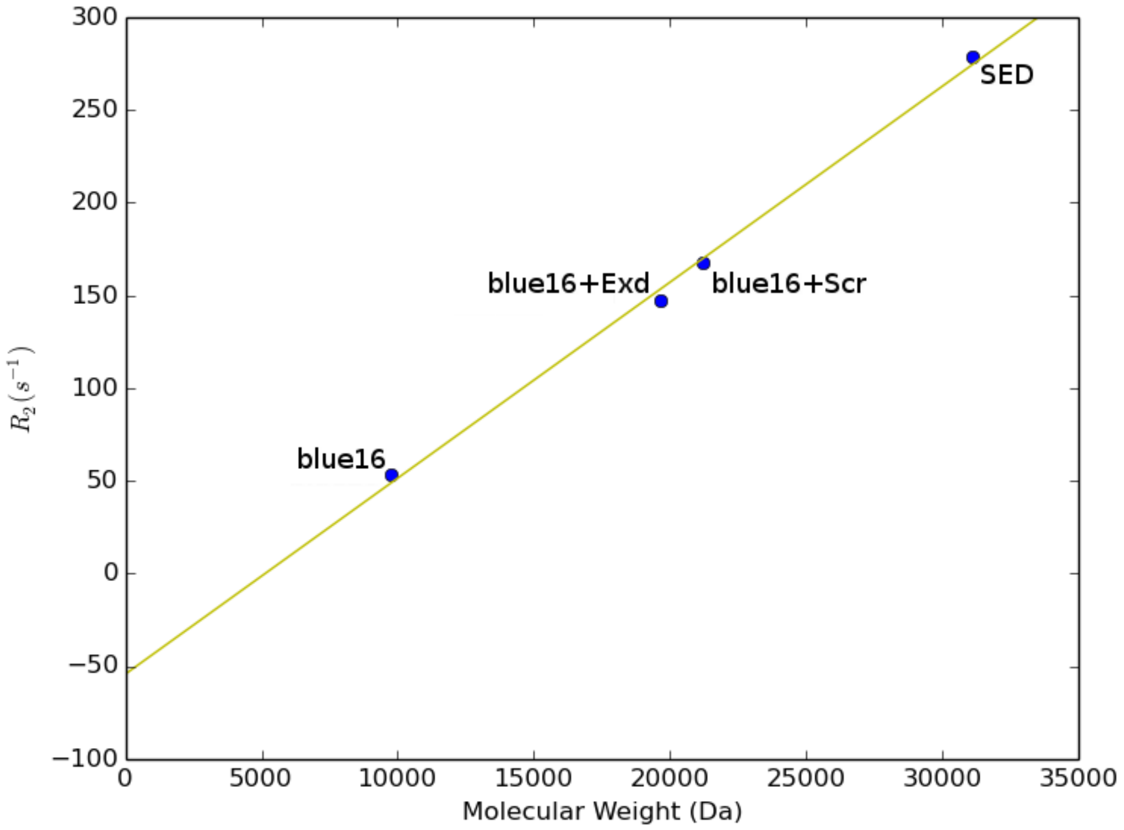


Figure 5.7: **Transverse relaxation rate versus expected molecular weight** Transverse relaxation rates R_2 versus expected molecular weights of the different complexes.

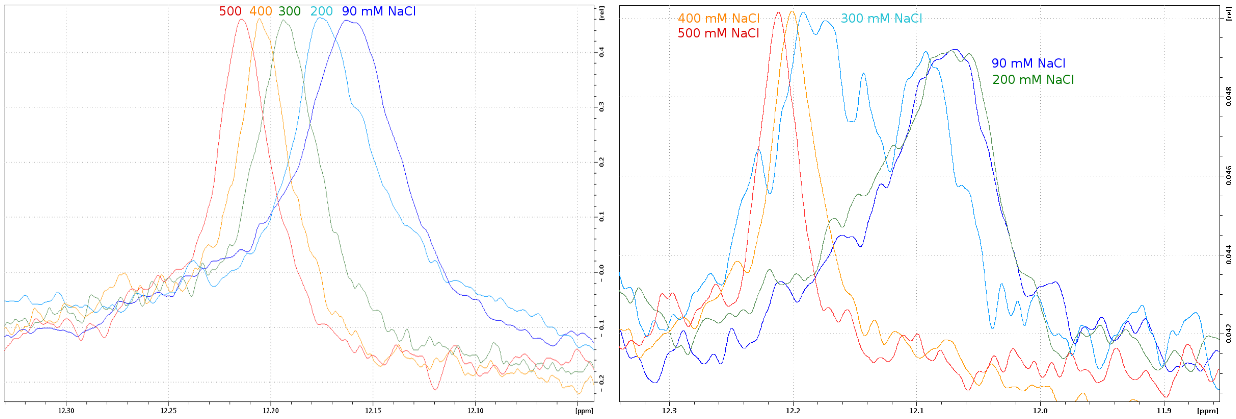


Figure 5.8: **Ternary complex is in slow exchange with free DNA**
The left panel shows Guanine 5 in the case of blue16-ExdHD. Adding salt gradually sharpens the peak until the linewidth is very close to that of free DNA. In the case of the ternary SED complex (right panel), adding salt doesn't seem to affect the peak shape much until the salt concentration reaches 300 mM, when a second, sharper peak appears downfield of the initial peak. The initial peak then disappears at higher salt concentration. At 300 mM there seem to be more than just two peaks suggesting a mix of spectra from different intermediate complexes, such as free DNA, DNA bound to only one of the two homeodomains and the ternary complex.

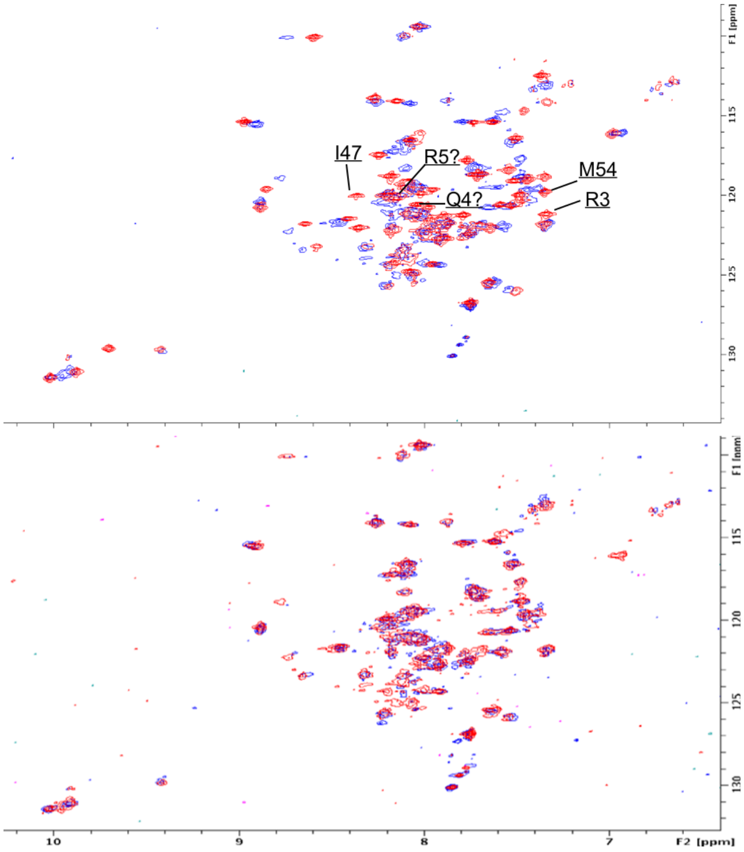


Figure 5.9: **2D backbone amide TROSY of Scr before and after adding DNA and ExdHD**

The top panel shows ScrND in red before and blue after addition of blue16. A few assignments of peaks important for binding have been transferred from Dr. O'Connell's previous assignments [23]. A number of peaks undergo shifts, including several that we know are important for binding. The bottom panel shows ScrND+blue16 before (red) and after (blue) addition of ExdHD. Few peaks shift, which might mean that Scr undergoes very little changes upon binding of ExdHD. Alternatively binding might not be seen. As mentioned in the text, this sample was missing $MgCl_2$ and was somewhat aged, and thus Scr could have possibly lost its N-terminal arm and linker region, thereby losing its ability to bind ExdHD. The fact that the linewidths do not seem to increase much upon addition of Exd, further suggests that indeed no ternary complex is formed.

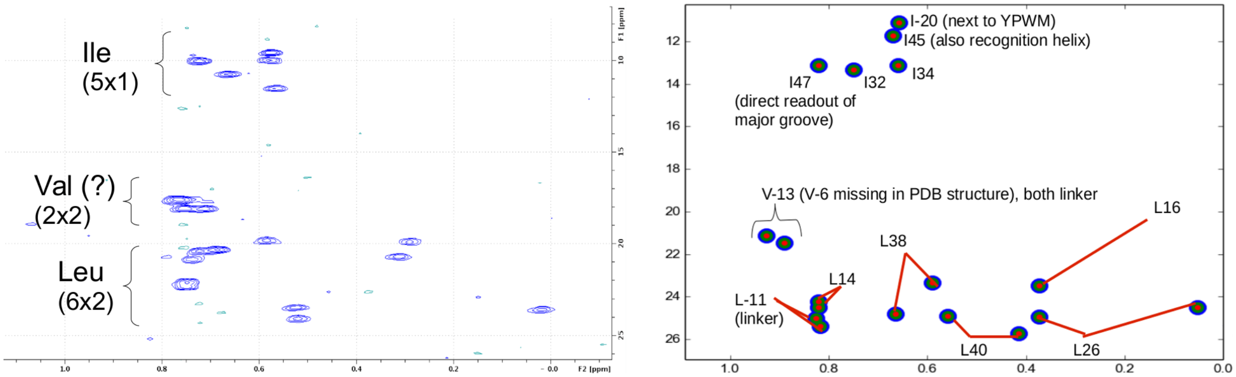


Figure 5.10: **ILV ^{13}C-methyl labeled Scr homeodomain and ShiftX2 predictions** The left panel shows a methyl TROSY spectrum for Scr collected on our in house 600 MHz spectrometer and a tentative assignment based on the methyl shifts predicted by ShiftX2 [122]. The panel shows a 2D plot of the shift predicted by ShiftX2. As can be seen the overall distribution of peaks maps pretty well to the actual spectrum making us confident to at least assign residue types for most peaks. For a real assignment, experiments have to be conducted that transfer magnetization from the methyl to the backbone to correlate the methyl shifts to the backbone shifts. Val-6 was absent from the structure and has thus no predicted peaks. Leu16 only had a hydrogen prediction for one of its two methyl groups. The missing methyl group has a predicted carbon shift of 24.9 ppm.

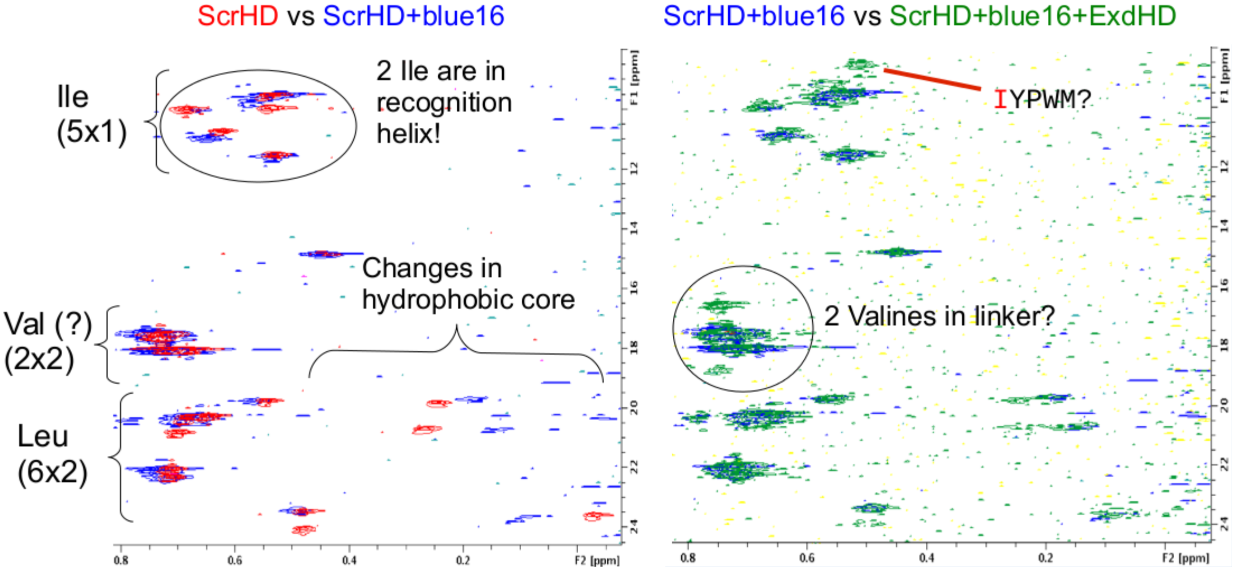


Figure 5.11: **Methyl-TROSY of ScrHD before and after adding DNA and ExdHD** The left panel shows the methyl region of ScrHD before (red) and after (blue) adding blue16. As can be seen there are a few peak shifts in the Isoleucine region, two of which lie in helix 3, which directly interacts with the DNA. Further shifts happen in the Leucine region which suggest a general rearrangement of the hydrophobic core of the protein. The right panel shows blue16-ScrHD before (blue) and after (red) adding ExdHD. Further changes happen in the Isoleucine region, which could be due to the Isoleucine next to the YPWM motif which is known to interact with Exd. Furthermore, new peaks appear in the region that we believe to be Valine peaks. This would make sense because both Valines in Scr lie in the linker region which will presumably be stabilized upon interaction with Exd. A few new peaks also appear in the Leucine region, but we believe that due to the long duration of the experiment the spectrum is a mix of free ScrHD, ScrHD-blue16 and the ternary complex because of degradation of Scr. Indeed collecting a 1D imino spectrum after the experiment indicates that at the end of the experiment no ternary complex is left.

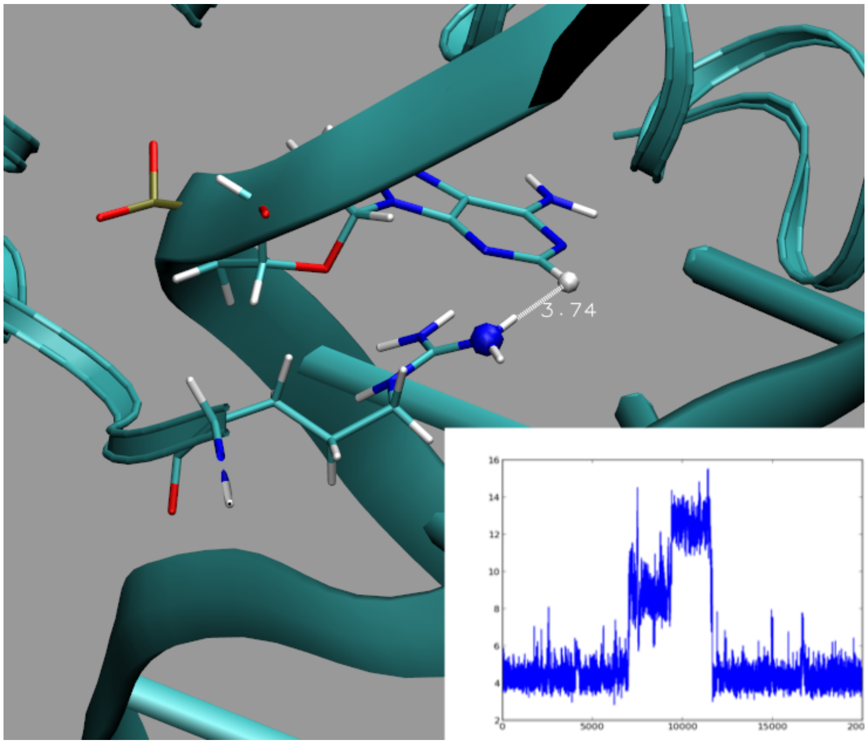


Figure 5.12: **The distance between Arg3 and A13 serves as a measure of insertion** The distance between the guanidinium group of Arg3 and the purine ring of Adenine 13 servers as a measure for Arginine insertion into the minor groove. Three states can be observed for wtScr on fkh: 1) inserted, 2) side chain bending back to interact with phosphate backbone and 3) the entire amino acid coming out of the minor groove. For the majority of the trajectory Arg3 stays inserted into the minor groove (state 1).

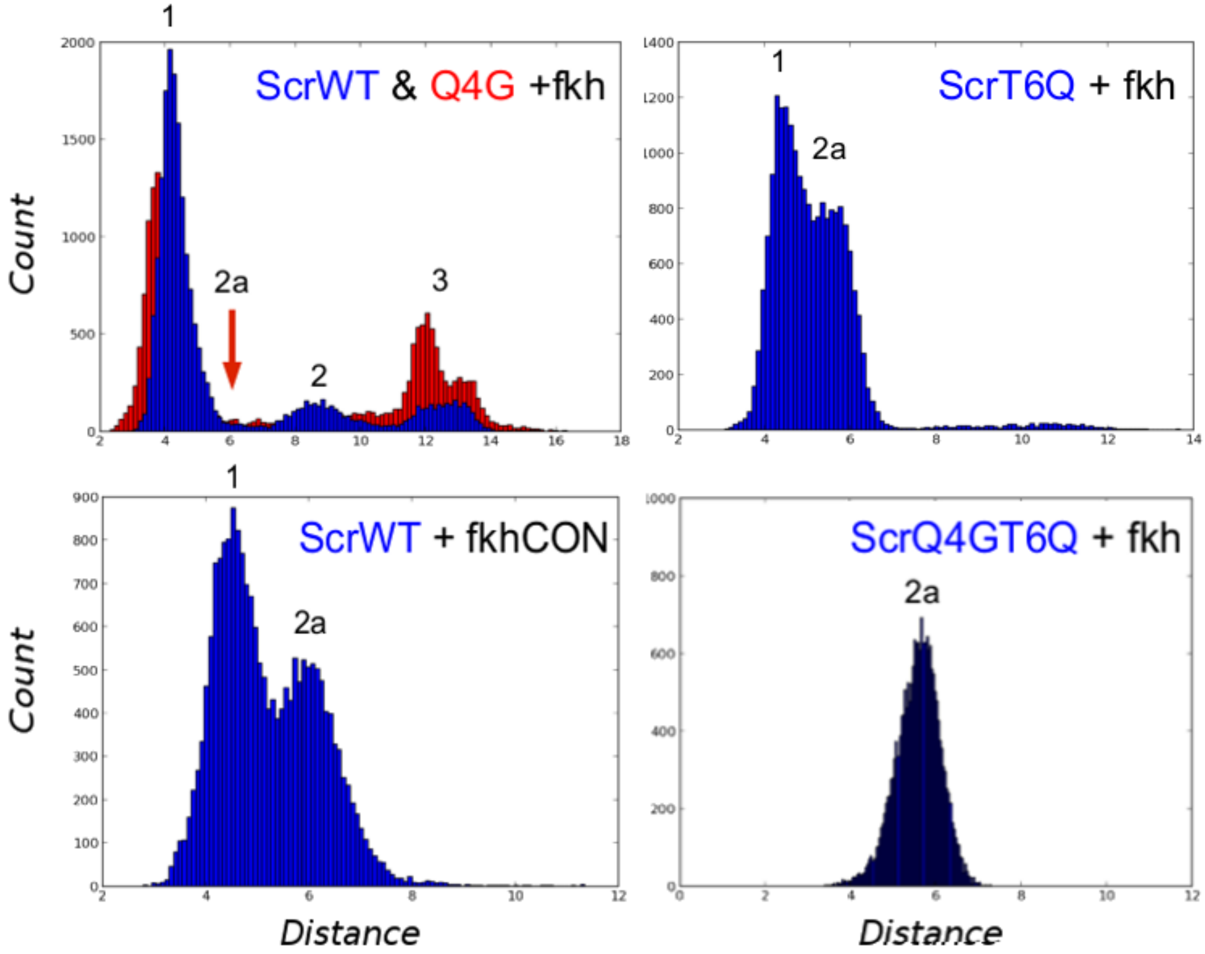


Figure 5.13: **Histograms of the Arg3-A13 distance reveal different populations for the different simulations**

The histogram of the distance of Arginine 3 and Adenine 13 reveals some interesting features. First, wtScr in combination with the canonical target fkh sees the clearest preference for the canonical, "inserted" state 1. ScrQ4G sees a slight increase of state 3 at the expense of state 1. Neither of them populate state 2a at all. Interestingly both for fkhCON as well as the the trajectories containing the T6Q mutation see an enormous increase in the population of state 2a. This is most pronounced for the double mutant, which exclusively populates this state.

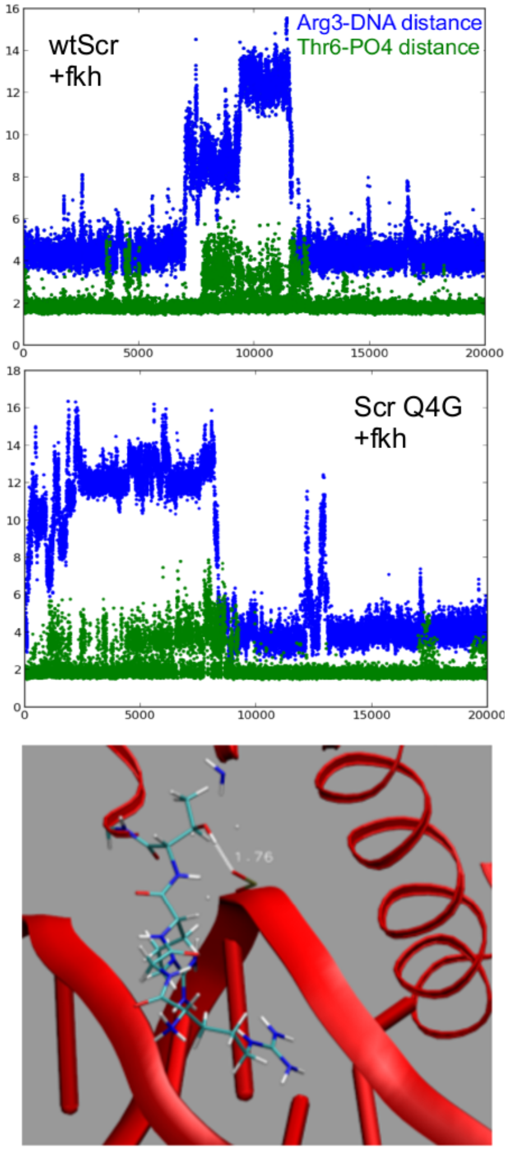


Figure 5.14: **Strong negative correlation between Arg3 insertion and breakage of Thr6 hydrogen bond**

The trajectories for fkh with wtScr and Scr Q4G show a strong correlation between Arg3 leaving its canonical inserted state 1 and the breakage of the hydrogen bond of Threonine 6 with the phosphate backbone of the DNA.

Chapter 6

Study of an AbdB-Cofactor-DNA complex by X-ray crystallography

6.1 Introduction

This chapter is a continuation of work done by Dr. Nithya Baburajendran and Anna Kaczynska of the laboratories of Richard Mann, Barry Honig and Lawrence Shapiro.

The overarching question we tried to answer in this chapter is the same one as in the previous chapter (chapter 5). How do Hox proteins achieve specificity, that is why do different paralogs prefer different target DNA sequences *in vivo*, even though they are highly conserved and bind to very similar sequences *in vitro* [112, 116, 123, 124]. Hox proteins, as well as other transcription factors and DNA binding proteins, must select a small number of target sites from a vast pool of possible binding sites available in a typical eukaryotic genome. For a more on Hox proteins please refer to the introduction in chapter 5. As specificity we define differences in affinities between Hox paralogs to the same DNA sequence. A DNA sequence can be low affinity for all Hox paralogs and still show specificity to a certain Hox paralog. Because low affinity binding sequences are hard to identify, we usually talk about high affinity sequences throughout the text, referring to an interaction as specific when there is a preference of the sequence for one Hox paralog over another, or sometimes when there is a preference of a Hox paralog for some sequence over another.

Joshi et al [49] showed that the *Drosophila* Hox protein Scr can bind to its *in vivo* target fkh, while other fly Hox paralogs fail to do so, and that it does so by “reading” a local minimum in the minor groove width with a conserved Arginine residue (Arg3). The group solved two crystal structures of the homeodomains of Scr and its cofactor Exd, one bound to fkh and one bound to a mutated concensus sequence fkhCON, which can be bound by several Hox paralogs (see Figure 5.1). The structures show that Arg3 of Scr inserts into a local minimum in the minor groove when bound to fkh but does not insert this residue into the minor groove when bound to fkhCON where no local width minimum exists. The hypothesis from the publication was that Scr, while binding similar sequences as other Hox proteins *in vitro* in the absence of Exd, interacts with the homeodomain of Exd through its YPWM motif (located on its N-terminal linker region), thus positioning the N-terminal arm (”NTA”, residues 1-9 of the homeodomain) along the minor groove allowing Arg3 to “read out” the local minimum in minor groove width.

In 2011, Slattery et al. developed a high throughput approach to characterizing binding affinities of all *Drosophila* Hox proteins in the presence and absence of its cofactor Exd to a large library of DNA sequences. In *Drosophila* more than 50 homeodomain proteins all prefer the binding sites TAATTG and TAATTA [125, 126]. They do however have specific target binding sites *in vivo* that are often not shared with other homeodomain proteins (one example is fkh, described above). One of the proposed solutions to this dilemma is a mechanism that they dubbed “latent specificity” by which a certain transcription factor (like Hox) reveals its specific DNA binding preferences only in the presence of its cofactor (like Exd). While this idea had been around for a while and differences in sequence preferences between Hox proteins and Hox-Exd complexes had been reported previously

[113, 124, 127, 128], Slattery et al. were the first to systematically analyze the influence of ternary complex formation had on specificity. They introduced an experimental/theoretical approach called SELEX-Seq that allowed them to calculate relative affinities of Hox and Hox-cofactor complexes to a large random library of DNA target sequences. They purified all height Hox proteins and performed gel shift assays with a DNA library of random 16mers flanked by sequences needed for PCR amplification in the presence of absence of Exd. Oligomers that showed enrichment in the gel shift assay were sequenced and used as a new, smaller pool of DNA sequences for a subsequent round of gel shift assays. This cycle was repeated for several rounds, allowing the team to calculate relative affinities of all sequences in the original library by analyzing round to round enrichment. The group was thus able to generate a relative affinity profile for all *Drosophila* Hox proteins with and without cofactors to $>10^9$ distinct 16mer binding sequences. The results strongly support the idea of latent specificity. All Exd-Hox heterodimers were found to prefer the sequence GAYNNAY (Y = T or C) and sequence preferences of the different paralogs were now strongly dependent on the identity of the paralog. Resulting heterodimer binding sequences were grouped by their 8mer core motif and classified according to a color scheme (red, blue, green, magenta, light blue, yellow, black, orange, light green and purple core motifs; see Figure 5.3 in previous chapter). The resulting preference profiles show clear affinity preference fingerprints for each Hox protein in the presence of Exd, while showing much less specificity (lack of preference) in the absence of Exd. Using Monte Carlo simulations they also calculated DNA shape predictions and thus DNA shape preferences for different Hox proteins. These shape predictions support the idea of shape readout from the earlier publication by Joshi et al. [49], that anterior Hox proteins prefer DNAs that contain a second minimum along the core motif, for example blue and green motifs (where, presumably, the residue at position 3 inserts into the minor groove), while posterior ones prefer sequences with no such minimum, for example red and magenta motifs. They implied that the minor groove width is a preformed feature of the DNA at that position and is simply being "read out" by the Hox proteins' N-terminal arm with the help of residues like Arginine 3.

A recent study by Abe et al. [118] confirms these ideas. The group mutated residues in the N-terminal arm of Scr that they deemed important for DNA shape recognition and repeated the SELEX-Seq experiments with these mutants. The results showed that mutating residues Arg3 and His-12 (homeodomain numbering) resulted in 12mer DNA sequences being selected with a less narrow minor groove at position A_9Y_{10} than for wtScr. Even more strikingly, when mutating positions 4 and 6 of Scr to the respective residues found in Antp and Ubx (Q4G and T6Q mutations), the preference for a minor groove minimum at this position was almost completely abolished and the binding preference fingerprint was almost perfectly turned into that of Antp. This underlines the importance for the identities of the residues in the NTA for shape readout and paralog specificity.

A few questions that remain unanswered by these publications are:

1) Why do other Hox paralogs that also have this Arginine at position 3 fail to bind these same sequences?

2) Why does failure to insert the Arginine result in a failure to bind the sequence. This is especially puzzling because Scr can clearly bind the consensus fkhCON sequence without inserting the Arginine.

3) Is the DNA shape predetermined by its sequence and independent of Hox and Exd binding? Is Arginine 3 insertion in the fkh structure a result of the narrow minor groove or the other way around?

4) Do the predicted minor groove shapes hold up when examined with X-ray crystallography and is there a clear correlation between actual minor groove width and Hox affinity?

5) Is a local minimum in the DNA sufficient for insertion of residue 3 into the minor groove? That is, will posterior Hox proteins also insert residue 3 into the local minimum

Table 6.1: **The four studied DNA motifs**

Oligomer name	Oligomer sequence
red14	GCATGATTTATGAC
blue14	GCATGATTAATGAC
magenta14	GCATGATTTACGAC
black14	GCATGATAAATGAC

The Exd core binding motif is highlighted and the variable Hox core binding motif is underlined.

of a sequence that is not a preferred binding site? This question is related to question 3.

6) If the shape of the DNA dictates the conformation of the NTA for anterior Hox proteins like Scr, does this apply to other Hox proteins as well, in particular posterior ones? In other words, are all Hox proteins equipped to recognize DNA shape or only some?

We addressed questions 1 and 2 in the previous chapter (chapter 5) and wanted to use X-ray crystallography to address questions 3-6, as well as collect further data to improve our answers for questions 1 and 2. To this end, we crystallized the homeodomain of most posterior *Drosophila* Hox protein AbdB in complex with the homeodomain of Exd, bound to four different DNA sequences, namely members of the red, blue, magenta and black families as described above. The four sequences are listed in table 6.1 and will henceforth simply be called red14 or red, blue14 or blue, magenta14 or magenta, black14 or black.

Dr. Baburajendran recorded X-Ray diffraction for crystals AbdB in complex with Exd and the red and blue oligomers. Guided by her screens, we designed new crystal screens for AbdB and Exd with the black and magenta sequences. AbdB is the most posterior of the *Drosophila* Hox proteins, as well as the least conserved one. It obeys the "posterior prevalence" (sometimes "posterior dominance") phenomenon [129–131], meaning it can outcompete anterior Hox proteins for many binding sites, possibly in part due to its short linker region, the shortest one among the *Drosophila* Hox proteins. AbdB's short linker also makes it one of the best suited Hox proteins for crystallography, in particular when complexed with DNA and Exd, which presumably immobilizes the linker further. The vertebrate AbdB ortholog HoxA9 has previously been solved in complex with Pbx1 using X-ray crystallography by LaRonde-LeBlanc & Wolberger [129] (PDB: 1PUF) and the AbdB ortholog HoxA13 has been solved in its bound and unbound form using solution NMR by Zhang et al. [132] (PDB: 2L7Z & 2LD5).

6.2 Different target DNAs identified by SELEX-Seq

6.3 Crystallization

Initial 96-well crystallization screens were designed to replicate conditions found by Nithya Baburajendran and Anna Kaczynska's screens. The conditions she found for diffraction of her crystals of AbdB-Exd14 with red14 and blue14 (both blunt ended), were 200 mM $MgCl_2$, pH 5.8 (100 mM TRIS), 17.5% w/v PEG 4000 and 5% v/v glycerol for red14 (diffraction to 2.44Å) and 200 mM $MgCl_2$, pH 5.8 (100 mM TRIS), 17.5% w/v PEG 3350 and 2.5% v/v glycerol for blue14. It should be noted that her complex formation buffer was different from ours. She used 200 mM NaCl, 10 mM Tris pH 8.0, 2 mM TCEP. The main difference to our buffer is the absence of magnesium chloride or any other divalent cations. As described in the Materials and Methods chapter 2 and chapter 5, magnesium is important for complex formation. When trying to replicate her protocol (not using magnesium chloride), we encountered precipitation as expected.

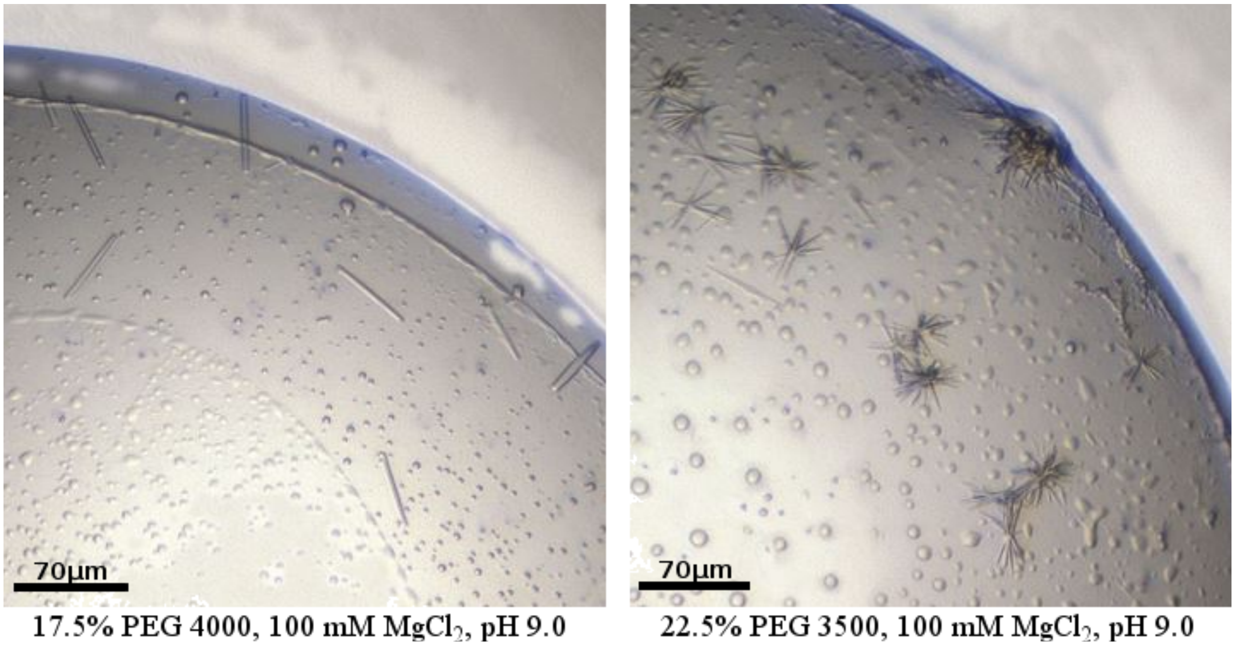
Our first screen replicated Dr. Baburajendran's complex formation conditions for AbdB, Exd14 and black14 and varied PEG 3350 and PEG 4000 from 15 to 22.5 % and $MgCl_2$ from 150 to 300 mM, while screening pH 5.8, 7.0 and 8.5. A few drops contained interesting aggregates but nothing that seemed worth pursuing further.

We presumed that adding $MgCl_2$ from the start should increase complex stability and the actual complex concentration in the drops by avoiding precipitation, and thus improve crystallization behavior. For all of the following crystallization screens we thus decided to use the conditions listed in table 2.1 of chapter 5 (”Crystallization buffer”).

Our next set of screens varied PEG 3350 and PEG 4000 from 15 to 22.5 % and $MgCl_2$ from 100 to 300 mM, while screening pH 5.8, 7.0 and 8.1 and 9.0. Some interesting crystalline structures (but with round edges) were found at the higher $MgCl_2$ concentrations for both complexes, but they showed no fluorescence under UV light and were thus presumed to be salt crystals (presumably magnesium chloride, in spite of the round edges). At the lower end of the $MgCl_2$ concentration spectrum we found more interesting rod and spike like structures for both complexes that showed mild fluorescence under UV light. We set up 4 screens total, two for each type of DNA (magenta14 and black14), one with the shorter Exd14 construct and one with the longer Exd320 construct. The Exd homeodomain ends at residue 300 according to the full length numbering, whereas Exd14 extends to residue 310 and Exd320 to residue 320. The shorter Exd14 construct resulted in better initial hits, and was therefore used for all of the subsequent screens.

For magenta14, initial hits were seen at 100 mM $MgCl_2$ and the higher end of the pH (7.0-9.0) and PEG spectrum (mostly at 22.5%) (Figure 6.1). Given that the buffer of complex formation contained 50 mM $MgCl_2$ (as opposed to Dr. Baburajendran’s earlier screens) it seemed reasonable that hits were seen at lower magnesium concentrations than she had found previously.

Follow-up screens used a lower range of $MgCl_2$ concentrations for both complexes (0-150 mM), while increasing the PEG concentration range slightly (17 - 26 % w/v). These screens showed many hits for the “black” complex, almost entirely at the conditions without any $MgCl_2$ (see Figure 6.2), while the “magenta” complex showed a few hits in the mid-range



17.5% PEG 4000, 100 mM $MgCl_2$, pH 9.0

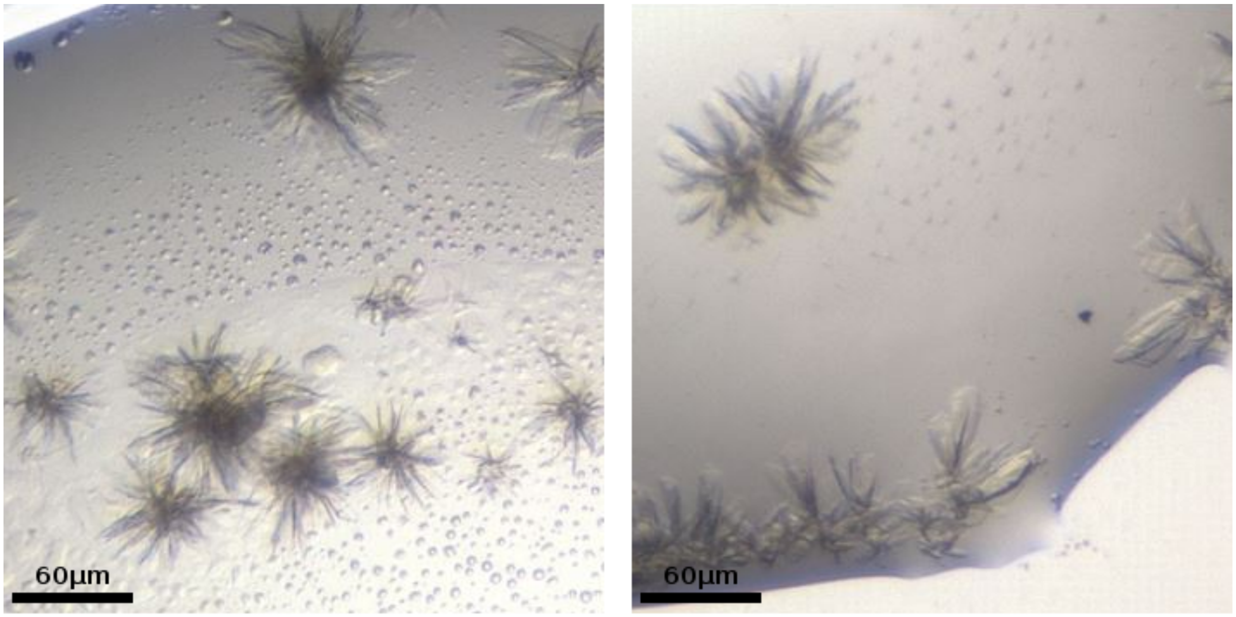
22.5% PEG 3500, 100 mM $MgCl_2$, pH 9.0

Figure 6.1: **Early crystal hits for magenta14**
Some early crystals for AbdB-Exd14-magenta14 and the conditions at which they formed.

magnesium concentrations (50 and 100 mM mostly, see Figure 6.1)). Both showed hits over a broader range of pH, but the black complex seemed to prefer lower pH and the magenta one higher pH. The best hits for these screens were obtained for the black complex at pH 7.0, 0 mM $MgCl_2$ at 17 and 20% PEG 3350. These two conditions show very nice clear blades with sharp edges growing out from a star like structure (Figures 6.2 and 6.3).

Several 24-well optimization screens were set up for the black complex around those conditions but crystals looked very poor in comparison to the crystals seen in the 96-well plates. We thus decided to optimize further on 96-well plates and pick crystals directly from those.

Several more crystallization screens on 96-well plates were set up for both black and magenta crystals, narrowing down the best conditions for each complex, by varying primarily magnesium concentration and pH, as well as the ratio of DNA to protein (which turned out



17% PEG 4000, 0 mM $MgCl_2$, pH 8.1 23% PEG 3350, 0 mM $MgCl_2$, pH 8.1

Figure 6.2: **Early crystal hits for black14**
Some early crystals for AbdB-Exd14-black14 and the conditions at which they formed.

to be less important than the buffer conditions). As the early screens suggested, the black complex preferred very low $MgCl_2$ concentrations (0-20 mM) and lower pH (5.3-7.0), while the magenta complex preferred higher concentrations (70-90 mM) and a pH of around 9.0.

The crystals that finally ended up diffracting best and that were used to solve the structures were 25% PEG 3350, 0mM $MgCl_2$, pH 5.3 (100 mM NaCitrate) for black14 and 22% PEG 3350, 90 mM MgCl2, pH 9.0 (100 mM TRIS) for magenta14.

All crystals that diffracted well were of the same blade like shape, measuring between 50 and 300 μm in length and 20-50 μm in width, but being very thin in the third dimension. 1 μl of cryo protecting solution (well solution plus 30% v/v glycerol) was added on top of the 200 nl drops and crystals harvested and mounted on nylon loops before being flash frozen in liquid nitrogen.

Crystals of other shapes, both from 96-well plates as well as 24-well plates were sent

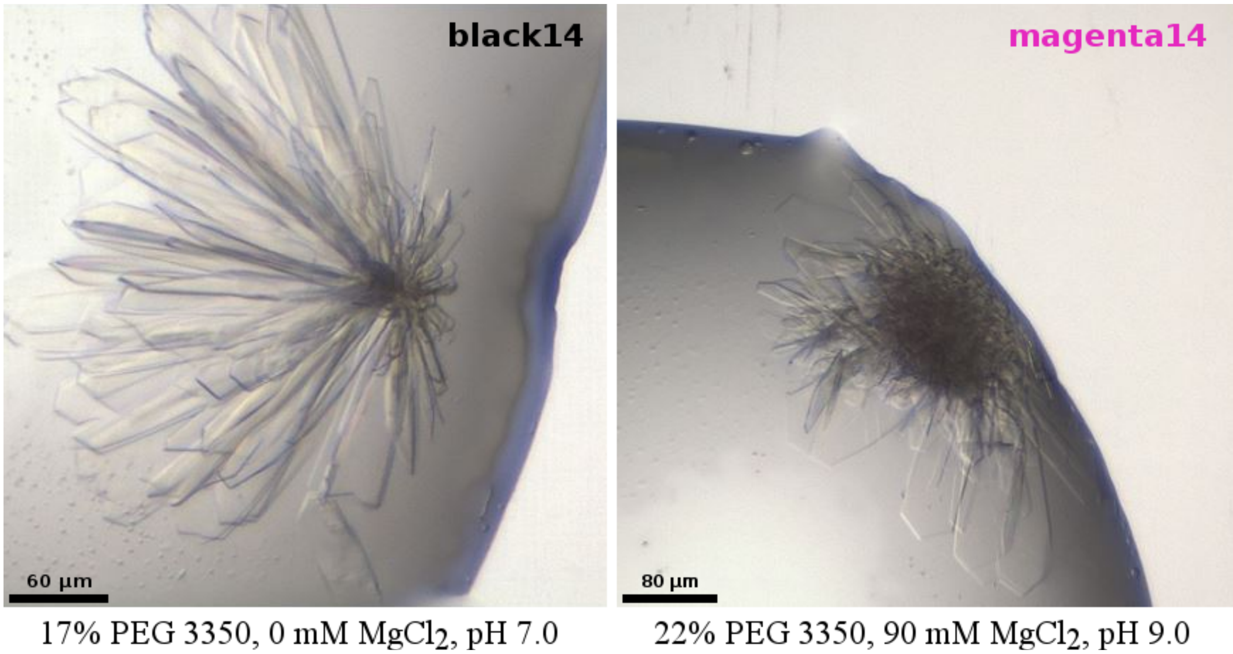

black14
60 µm
17% PEG 3350, 0 mM MgCl2, pH 7.0
magenta14
80 µm
22% PEG 3350, 90 mM MgCl2, pH 9.0

Figure 6.3: **Optimized crystals for black14 and magenta14**
Optimized crystals in 96-well trays for AbdB-Exd14-black14 and AbdB-Exd14-magenta14. The blade shaped crystals growing out from a star-shaped structure were characteristic of black14 crystals (left panel). The blade shaped crystals growing out from a bushy precipitate structure were characteristic of magenta14 crystals (right panel). Most of the blade shaped crystals diffracted very well to about 3Å. The blades were broken off from the cluster and mounted individually on loops for diffraction experiments.

to the synchrotron for diffraction experiments but showed inferior diffraction to the blade shaped ones (weak diffraction, higher mosaicity, lower resolution).

The blade shaped crystals picked from the 96 well plates proved to diffract much better (low mosaicity, diffraction to about 3Å) but did often not show diffraction at all orientations of the crystal, such that it was impossible to collect full datasets. The best datasets turned out to be collected from blades whose long axis was aligned with the long axis of the loop and that were thus rotated around their long axis. While it is possible that this is coincidence it seems plausible that using the crystals' long axis as rotation axis yields better results, if the crystal lattice has different qualities along different axes of the crystal.

Diffraction patterns were collected at the Advanced Photon Source at Argonne National Laboratory (Argonne, Illinois, USA) on beamline ID-24E. 200 images at 1° angle increments were collected. Mosaicities were relatively low with 0.703° and 0.268° and data was processed to 3.03 and 2.4 Å for magenta and black complexes, respectively. A summary of data collection and refinement statistics can be found in table 6.2, which also includes available statistics for the red and blue crystals, solved by Dr. Baburajendran and Anna Kaczynska.

Space groups were found to be C121 (C2) and P1 for magenta and black, respectively.

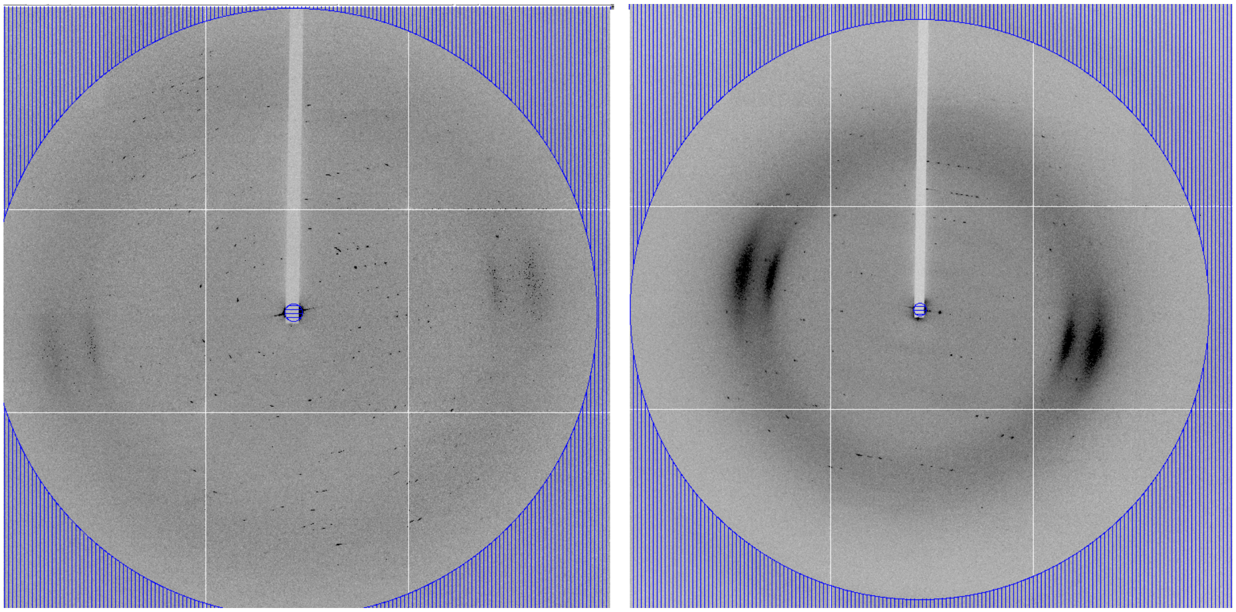


Figure 6.4: **Example diffraction patters for black14 and magenta14**
Example diffraction patters for AbdB-Exd14-magenta14 (left panel) and AbdB-Exd14-black14 (right panel). The blue circle on the left corresponds to the 3Å resolution limit and the circle on the right corresponds to 2.4Å.

Table 6.2: **Crystallographic table for all four structures**

	AbdB-Exd14 -red14	AbdB-Exd14-blue14	AbdB-Exd14-black14	AbdB-Exd14-magenta14
Data collection				
Space group	C2	C2	P1	C2
Cell dimensions				
a,b,c (Å)	77.06, 49.4, 95.19	77.76, 49.66, 96.8	45.44, 45.6, 66.86	77.33, 49.45, 95.1
α, β, γ (°)	90, 109.3, 90	90, 109.0, 90	99.0, 100.4, 114.2	90, 109.2, 90
R_{sym}*	N/A†	N/A†	0.169 (0.706)	0.237 (0.629)
I/σ_I*	N/A†	N/A†	6.8 (2.0)	6.0 (1.7)
Completeness* (%)	0.97	0.94	93.1 (92.6)	0.99 (1.00)
Multiplicity*	N/A†	N/A†	2.0 (2.0)	3.6 (3.5)
Refinement				
Resolution (Å)	29.94 - 2.443	38.76 - 2.897	63.55 - 2.4	44.91 - 3.03
Unique reflections	12408 (980)	7521 (653)	16923 (1662)	6678 (624)
R_{work}/R_{free}	0.234/0.255	0.253/0.274	0.206/0.259	0.248/0.285
Number of atoms	1746	1610	2918	1722
Protein	1135	1026	1646	1094
DNA	568	568	1139	571
Water‡	43	16	133	57
B-factors	46.8	70.2	38.6	58.4
Protein	51.0	74.0	39.4	61.0
DNA	39.4	63.9	38.0	54.0
Water‡	33.3	51.1	33.8	35.2
R.m.s deviations				
Bond lengths (Å)	0.004	0.003	0.003	0.003
Bond angles (°)	0.64	0.53	0.54	0.45

* Values in parentheses are for the highest resolution shell.

† Data has not been provided by Dr. Baburajendran

‡ Ions were modeled as waters

6.4 Description of four crystal structures showing AbdB and Exd bound to four different target sites

This section will describe the two structures solved as part of this study (magenta and black) as well as the two structures solved by Dr. Baburajendra and Anna Kaczynska (red and blue), which we analyzed together to make a thorough comparison of AbdB and Exd bound to all four oligomers. The resolutions of the red and blue crystals are 2.44Å and 2.90Å , respectively.

The red and blue crystals are in the C2 space group just like the magenta crystal, all three having about the same unit cell dimensions (see table 6.2). The black complex was in the P1 space group and is an outlier in more than one way. The former three have a single DNA duplex with one homeodomain of AbdB and one homeodomain of Exd in the asymmetric unit - that is, one single ternary complex. The black complex on the other hand, the only one in P1, has an extra DNA duplex bound by a single AbdB homeodomain but no Exd homeodomain in the assymetric unit - that is, the asymmetric unit contains one ternary complex and an additional binary complex, which is missing Exd. The AbdB in this binary complex is recognizing an additional Hox binding site the the reverse complementary strand of the DNA that is missing an Exd binding site (*vide infra*).

Ignoring the extra binary AbdB-DNA complex in the black crystal, all four ternary complexes show the typical binding mode observed in other Hox-Exd-DNA structures [49, 108,117,129,133]. AbdB and Exd bind in head-to-tail fashion to opposite faces of the DNA, using overlapping binding sites, with their respective recognition helices (helix 3 of the homeodomain) lying in the major groove of the DNA, its side chains making direct contacts with the DNA nucleotides (see Figure 6.5). The protein backbones moieties of all four complexes superpose very well, with a C_α RMSD of < 1Å for any pair of homeodomains, when aligned by the DNA moieties.

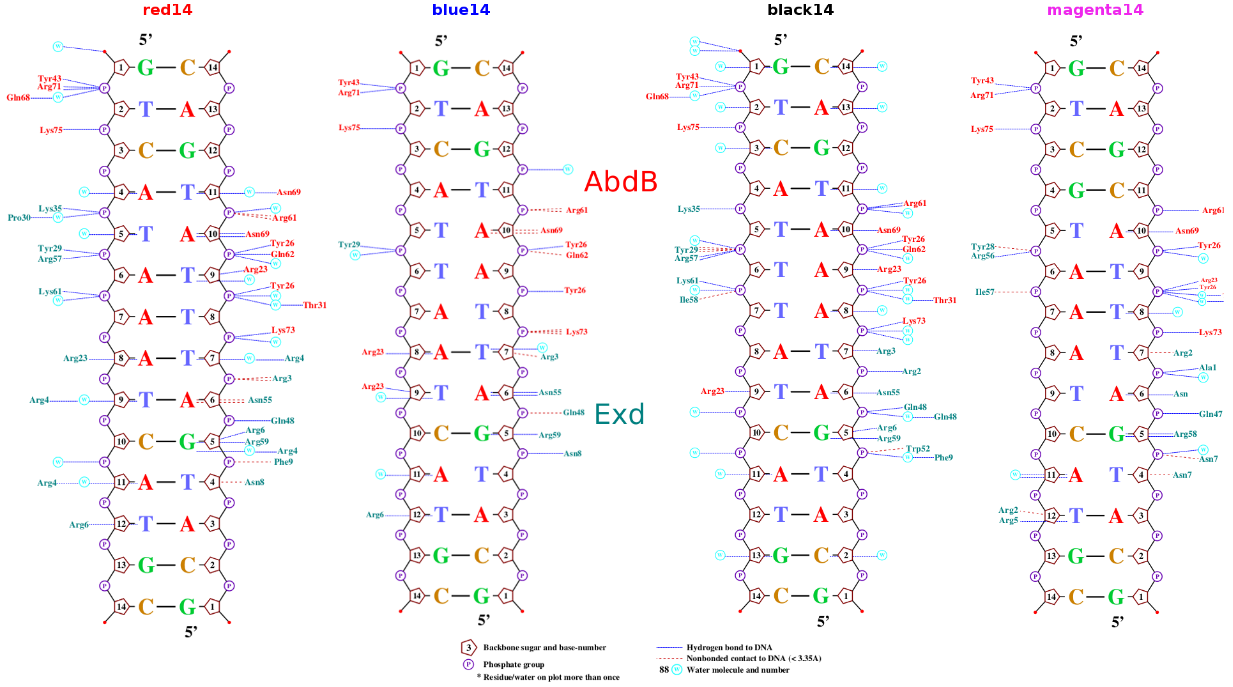

red14
blue14
black14
magenta14
5'
AbdB
Exd
Backbone sugar and base-number
Phosphate group
* Residue/water on plot more than once
Hydrogen bond to DNA
Nonbonded contact to DNA (< 3.35A)
88 Water molecule and number

Figure 6.5: **Protein-DNA interaction map**
Protein-DNA interactions for all four ternary complexes. Plots were made using NucPlot [134].

AbdB, like all Hox homeodomains, consists of three helices, an N-terminal arm (residues 1-9) and linker region N-terminal to the homeodomain that includes the hexapeptide, often called YPWM motif for its central consensus sequence for 6 of the 8 *Drosophila* Hox proteins. In the case of AbdB, which is the most divergent *Drosophila* Hox protein from a sequence perspective, the YPWM motif is replaced by an HEWT motif, with the conserved tryptophan being responsible for interaction with the Exd homeodomain by inserting into a hydrophobic pocket. AbdB also has the shortest linker region of all the *Drosophila* paralogs (three residues between homeodomain and hexapeptide, compared to ranging from 8 to 109 residues for the other *Drosophila* paralogs). The hexapeptide (LHEWTG, in the case of AbdB), and in particular the tryptophan, still binds to the same hydrophobic pocket as for the other Hox proteins, formed by the so called TALE motif of the Exd homeodomain (three amino acid loop extension, inserted between helices 1 and 2 of Exd) [49, 108, 117, 129, 133], forming a hydrophobic pocket with the C-terminal end of helix 3 (see Figure 6.7). While some density for the HEWT motif is present in its binding pocket on Exd for all four ternary complexes, occupancies vary greatly between the structures (*vide infra*).

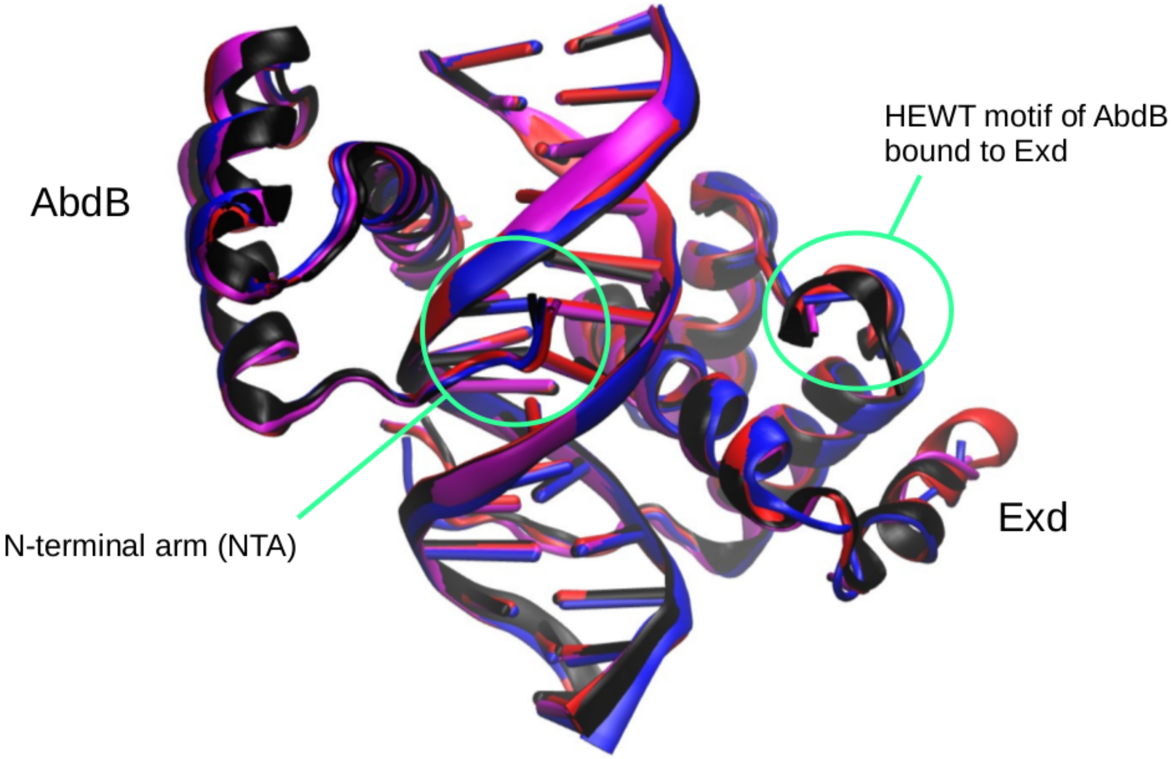


Figure 6.6: **Superposition of all four ternary complexes**
All four ternary complexes have been superposed by minimizing RMS deviation of one of the two DNA strands. The color code is according to the color of the DNA motif. The AbdB homeodomain can be seen bound to the DNA from the left, the Exd homeodomain from the right. The HEWT motif of AbdB is bound to Exd by inserting its tryptophan into a hydrophobic binding pocket. The N-terminal arm lies across the minor groove of the DNA. Most of the linker region between the NTA and the HEWT motif is missing and thus flexible.

Most interactions of the proteins with the DNA are conserved between the structures (see Figure 6.5). One of the biggest differences is the number of water mediated contacts. It is difficult to make an unbiased assessment of these contacts, because the positions of

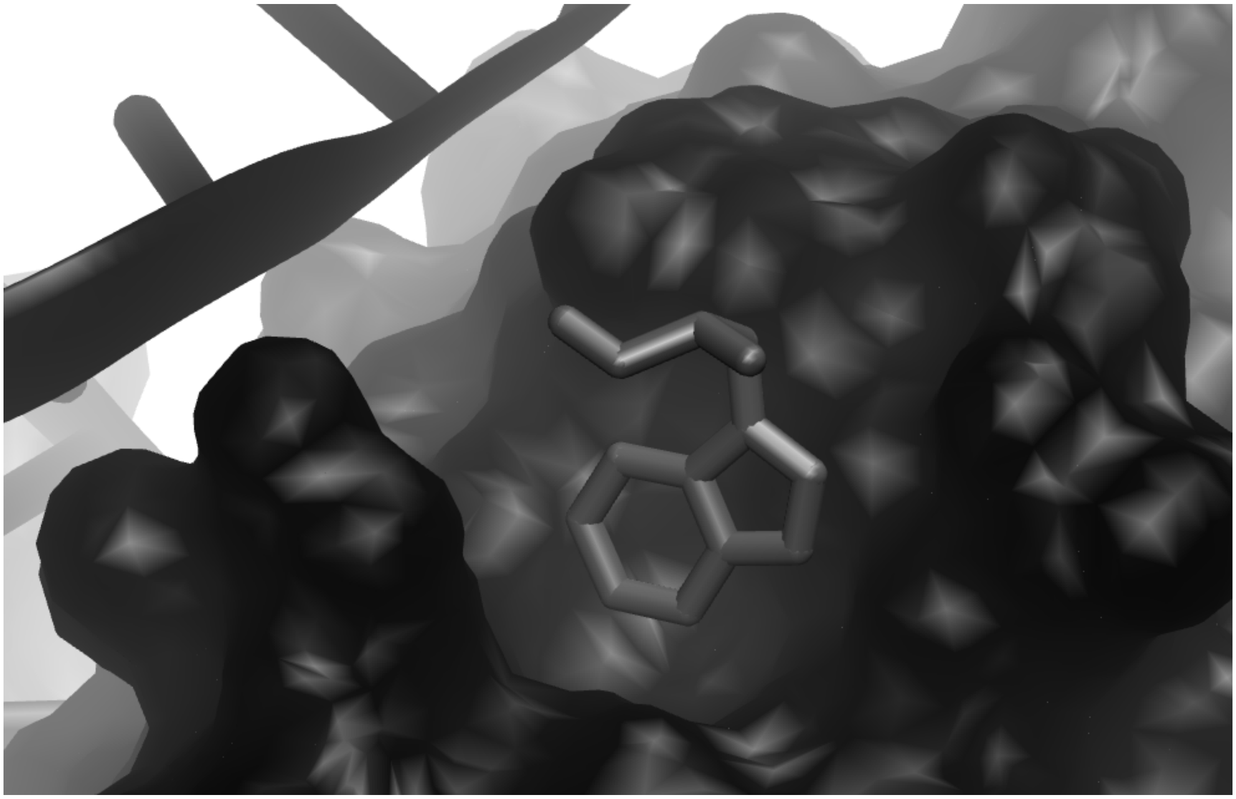


Figure 6.7: **Tryptophan of HEWT motif bound to hydrophobic pocket on Exd** The tryptophan of the HEWT motif of the ternary complex in the black crystal, bound to the hydrophobic pocket on the Exd homeodomain. The hydrophobic pocket is formed by the TALE motif and helix 3.

waters have relatively high uncertainties in general, and the ability to build them into the model is greatly dependent on the resolution and quality of the crystal structure. Water mediated contacts (and even more importantly a lack thereof in some structures) should thus be interpreted conservatively. The recognition helices of both homeodomains lie across the major groove of the DNA, while the N-terminal arms (residues 3-9) lie across the minor groove on the opposing face of the DNA, similar to other Homeodomain-DNA structures in the literature [49,108,112,117,129,132,133,135,136]. In the same way most contacts of the homeodomains in the major groove are identical to the ones described in the literature. The main differences in DNA contacts lie within the Hox NTA region and will be described below.

6.5 An extra binary complex binding in reverse fashion: AbdB-black14

As mentioned above the asymmetric unit of the “black crystal”, unlike the other three crystals, contains more than just one ternary complex. In addition to the canonical ternary complex, there is an additional binary complex of black14 DNA bound by only the AbdB homeodomain. Interestingly, this homeodomain does not bind to the same region of the DNA as in the ternary complex of the same DNA. Rather it sits on the opposite face of the DNA, its recognition helix occupying much of the major groove that would otherwise be occupied by Exd, while its NTA lies across the same minor groove as the forward binding AbdB, but in reverse fashion. Further inspection reveals that AbdB binds in reverse fashion to the same DNA, recognizing a Hox half binding-site in the reverse complementary strand of the black DNA. This half binding-site is missing an Exd binding site, but is otherwise very similar to the “red” core binding site. The “red” core sequence (as defined in [120]) is TGATTTATGA. The reverse complement of the black sequence used herein is GTCATTTATCATGC (with the 10mer core that resembles the red core motif underlined). Indeed, when analyzing the shape profile of the black DNA in reverse fashion, it strongly resembles that of the red DNA (*vide infra*).

This additional binary complex provides us with the opportunity to compare AbdB binding with and without its cofactor Exd. Additionally, because its recognition of a binding site in the reverse complementary strand precludes Exd binding and thus the formation of a ternary complex on this binding site, this opens up some room for speculation on the competition of overlapping binding sites, in particular competition of Hox-Exd and

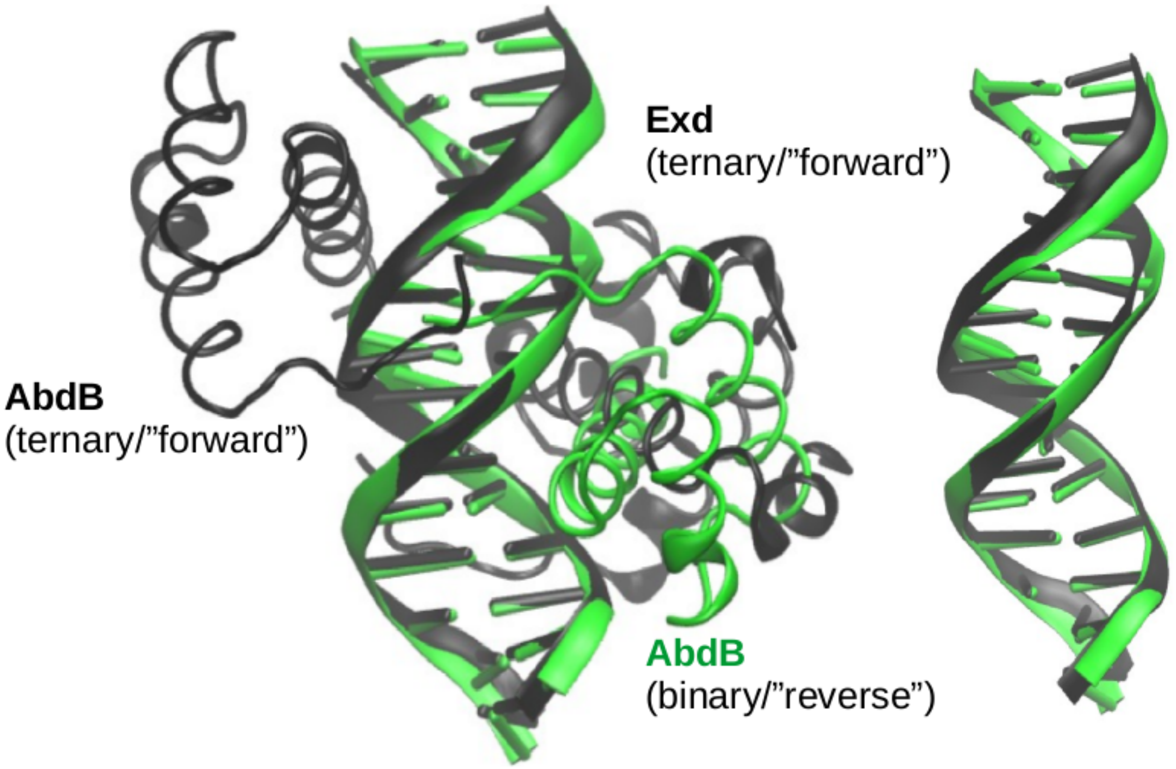


Figure 6.8: **Additional binary complex in black crystal**
A superposition by RMSD minimization of the black DNA for the binary (green) and ternary (black) complexes found in the black crystal. As can be seen, in the binary complex, AbdB does not bind in the same position as in the ternary complex but rather occupies the opposite face of the DNA, effectively recognizing an AbdB binding half site in the reverse complementary strand and blocking the Exd binding site in the canonical "forward" direction. The right panel shows a superposition of only the DNA duplexes to show differences in DNA shape caused by complex formation in two duplexes of identical sequences.

Hox only binding sites (see below for a discussion on this).

6.6 The AbdB N-terminal arm

For all five complexes described herein (four ternary and one binary complex) the NTA of AbdB lies along the minor groove of the DNA, although the occupancies and B factors of the NTAs of the different structures differ rather significantly (*vide infra*).

Arginine 5 of AbdB consistently inserts into the minor groove of the DNA for all structures while Lysines 3 and 4 do not (Figure 6.9). Both lysines have relatively weak density in all complexes. Lysine 4 extends its amino group towards the vicinity of a DNA phosphate, while lysine 3 extends its sidechain back towards the N-terminal end of the recognition helix of AbdB. Density of the Lys3 sidechain is weak for all structures but in the magenta and blue structures the density suggests a weak hydrogen bond with Gln44 and a water mediated contact with Thr41 of the AbdB homeodomain. Lys3 in the red crystal, while having density for its backbone in the same place as in the other ternary complexes, was modeled completely without a sidechain because no density was seen. Lys3 of AbdB in the extra binary complex of the black crystal could not be modeled at all, as even density for the backbone was missing. This suggests, as has been suggested before [49], that the interaction with Exd is an important contributor to the stabilization of the NTA.

The conformation of the NTA seen here corresponds exactly the its conformation in a crystal structure of the mammalian AbdB homolog HoxA9 in complex with the Exd homolog Pbx1 bound to DNA (PDB ID: 1PUF [129]). In contrast to the mammalian complex, we were not able to model the rest of the linker region between the HEWT motif and Lys3 (see Figure 6.9).

In contrast to the superposition with the mammalian complex, superpositions with the structures of the *Drosophila* paralog Scr bound to fkh and fkhCON (2R5Z and 2R5Y [49]) show differences in their NTA conformations (see Figures 6.10 and 6.11. In particular, when comparing residues 3 through 6 of our blue structure with 2R5Z (Scr bound to fkh, a specific *in vivo* target of Scr, related to the blue sequence), clear differences in backbone dihedrals can be seen for the peptide bonds betweeen residues 3 and 6 (Figure 6.11). Threonine 6 of Scr can be seen in hydrogen bonding distance with the DNA phosphate backbone (as discussed in the previous chapter, chapter 5), which appears to pull the NTA towards one

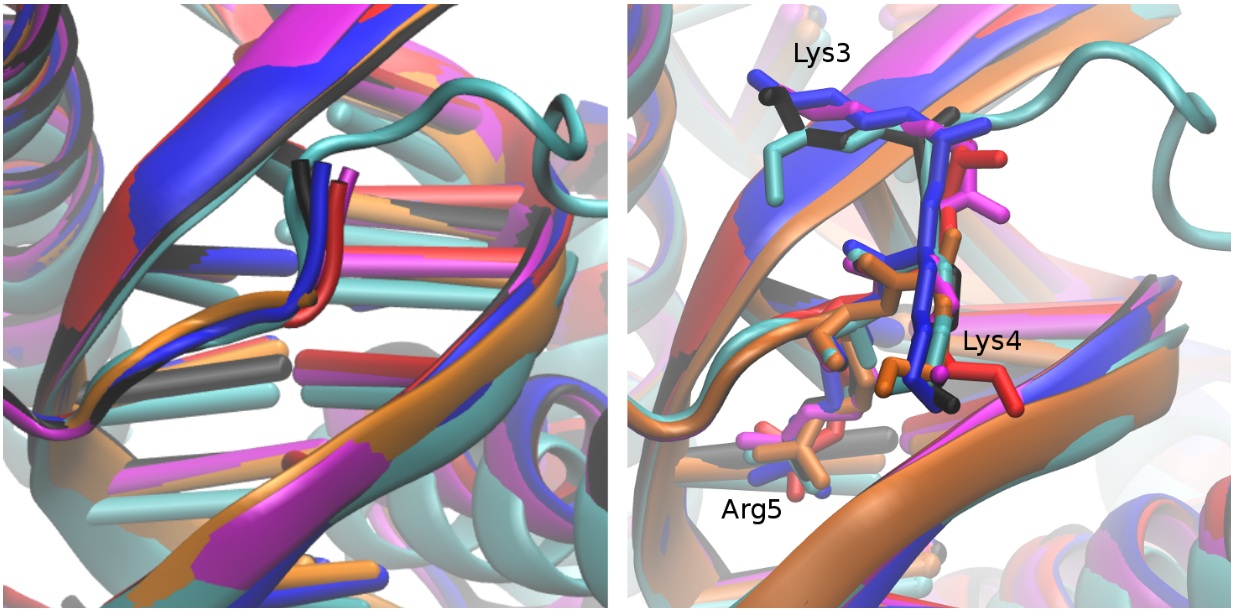


Figure 6.9: **The conformation of the N-terminal arm of all complexes is identical** A superposition of all five complexes described herein compared to the previously published ternary complex of the mammalian AbdB and Exd homologs HoxA9 and Pbx1 [129]. The left panel shows the NTA backbones only, while the right panel shows the sidechains of residues 3 to 5. The red, blue, magenta and black structures correspond to the ternary complexes of the same color. The orange structure corresponds to the inversely bound AbdB in the binary complex of the black crystal. the cyan structure corresponds to the HoxA9-Pbx-DNA structure described in the literature (1PUF, [129]). The NTA backbone conformation is clearly identical in all six structures and even the side chain orientations are mostly identical for the structures that showed clear enough density for them to be modeled. The mammalian complex had a much better resolution than any of our structures (1.9Å) and shows the entire NTA and linker region between Lys3 and the HEWT motif. Neither sidechain nor backbone of Lys3 could be built for AbdB in the case of the binary complex (orange), suggesting that its immobilization requires the interaction of the linker with Exd.

of the two DNA strands. Because of the constraints that are imposed by where Arg5 is inserted into the minor groove this seems to generate some torque on the backbone of the NTA which propagates to subsequent peptide bonds.

Gln4 of Scr on the other hand has no clear electron density and seems to simply stick out into the solvent, not interacting with the DNA in any way, allowing Arg3 to insert

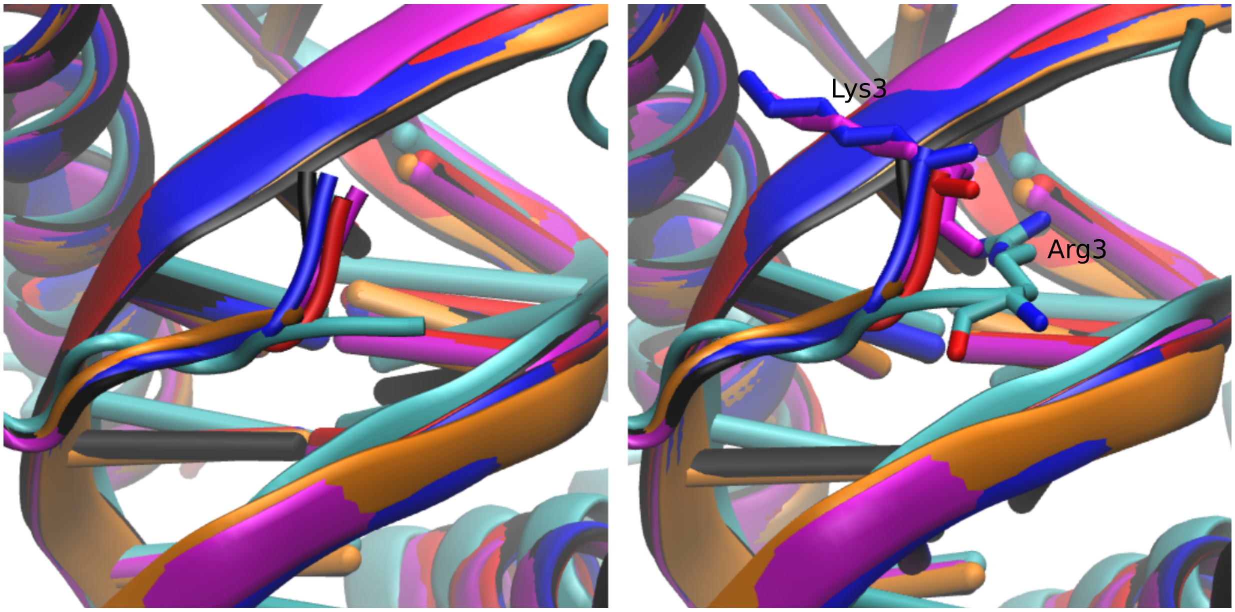


Figure 6.10: **AbdB N-terminal arm conformation differs from Scr**
A superposition of all five complexes described herein compared to the previously published ternary complex of the *Drosophila* Hox protein Scr in complex with its specific *in vivo* target fkh (related to the blue DNA) and Exd. The left panel shows the NTA backbones only, while the right panel shows the sidechains of residues 3. The red, blue, magenta and black structures correspond to the ternary complexes of the same color. The orange structure corresponds to the inversely bound AbdB in the binary complex of the black crystal and the cyan structure corresponds to the Scr-Exd-fkh structure described in the literature (2R5Z, [49]). The NTA backbone conformation of all AbdB structures is clearly different from Scr. The NTA of Scr is considerably twisted in comparison to the AbdB structures to accomodate for Arg3 inserting into the minor groove of the DNA.

into the minor groove. In the case of AbdB, Lysine 4 extends towards the backbone of the DNA. Since the ε-amino group of Lys4 is most likely protonated (pH 5.8), an ionic interaction with the phosphate backbone seems plausible. This puts enough strain on the NTA to not be able to rotate and allow Lys3 insert into the minor groove as seen for Scr. It has to be pointed out, that differences in dihedral angles C-terminal to residue 4, and thus presumably because of the lack of the hydrogen bond of residue 6 to the opposite strand of the DNA as seen for Scr, contributes to the interaction of Lys4 with the DNA, because its C_α atom already points towards the DNA backbone, whereas for Gln4 in the case of

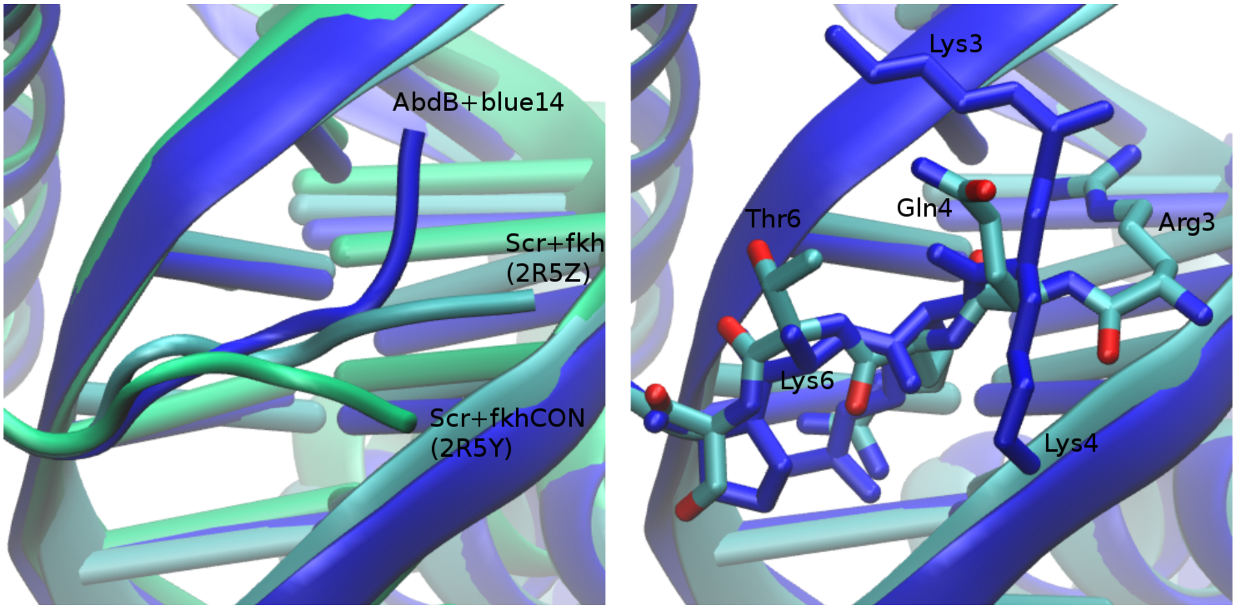


Figure 6.11: **Three different NTA conformations**
When superposing the backbone for our blue structure with both structures of Scr (one bound to its target blue-like structure - 2R5Z, one to a consensus red-like structure - 2R5Y), three very different conformations are seen. The left panel shows the backbones of the NTAs for our blue complex (blue), Scr bound to its *in vivo* target fkh (cyan) and bound to a consensus sequence (light green). The right panel shows a more detailed representation of the backbone and side chains of residues 3 through 6 for the blue structure and 2R5Z (Scr bound to fkh). Backbone dihedrals are strongly influenced by the sidechains interaction with the DNA, namely the interaction of Lys4 with the phosphate backbone in the case of AbdB, and the hydrogen bond of Thr6 with the DNA backbone as well as Arg3 insertion into the minor groove for Scr (see text for more details). While the backbone conformation is very similar for both Scr structures at positions 5 and 6, the differences are strong at positions 3 and 4, likely due to the insertion of Arg 3 into the minor groove putting additional restraints on the backbone conformation in the case of Scr bound to fkh (2R5Z).

Scr, the C_α atom seems to point straight into the solvent and not towards the DNA. The strain put onto the NTA by the interaction of Thr6 with the backbone can be seen more clearly even for Scr bound to fkhCON, where in the absence of the additional constraint of Arg3 inserting into the minor groove of the DNA, the NTA backbone gets even closer to the opposite strand of the DNA.

These findings clearly underline the importance of residues 4 and 6 for the conformation of the NTA and the insertion of residue 3 into the minor groove of the DNA (see also chapter 5). The slightly higher entropic cost of immobilizing a Lysine sidechain compared to an Arginine side chain [137] and the higher energetic cost to remove a Lysine from water [138] could also be a contributing factors. Importantly, the conformation of the NTA, at least in the case of AbdB, seems to be dictated not by the type of DNA it is bound to, but by the identity of the residues in the NTA. This could mean that the identity of the NTA residues make AbdB less sensitive to DNA shape than more anterior Hox proteins like Scr. This could be a factor contributing to the so called posterior prevalence phenomenon [129–131]. This also seems to answer question 6 we asked ourselves in the introduction of this chapter: AbdB does not seem to be as sensitive to DNA shape as Scr, at least judging by the conformation of the NTA. In the case of Scr, the interaction of Thr6 with the DNA phosphate seems to "twist" the NTA just enough for Arg3 to insert into the minor groove if a local minimum is present and thus recognizing its shape. It seems thus that not only the presence of an Arginine at position 3, but also the presence of a Threonine at position 6 (for the hydrogen bond) and the absence of a lysine at position 4 (which seems to want to interact with the phosphate backbone), contribute to Scr's ability to recognize DNA shape differences.

6.7 The AbdB HEWT motif and its interaction with Exd

There is some electron density for the HEWT motif for all four ternary complexes described here. The strength of the density, as well as the B factors of the tryptophans that we modeled into it, vary strongly, however. Both red and black ternary complexes have very clear electron density for the tryptophan and surrounding residues. In the case of the black ternary complex, we were able to model three residues N-terminal and two residues C-terminal to the tryptophan residue, meaning that we were able to build the entire hexapeptide LHEWTG. For the red complex, were were able to build only one residue on each side of the tryptophan, namely the sequence EWT. The same residues were built for the magenta complex, but where the density was well defined for the red complex, the magenta complex had very weak density even for the tryptophan, which we presume to be not as tightly bound and somewhat flexible, as opposed to the red and black ternary complexes. For the blue complex we only built the tryptophan and its preceding glutamate (sequence EW). Interestingly, it seemed impossible to build the same rotamer for the tryptophan as seen for the other structures (both described herein as well as in the literature). Instead, we ended up building a different rotamer, flipped by 180° which fits the density we found a little better (see Figure 6.12). Both the magenta and the blue crystals have high B factors for the tryptophans and R values were slightly higher when modeling them into the density than without. As a matter of fact, after building the tryptophan for the magenta complex, some negative difference electron density appeared for the tryptophan side chain (but not its backbone) suggesting that the modeled rotamer does not have full occupancy, presumably occupying the rotamer seen in the blue complex in part of the crystal. For both magenta and blue it seems that the HEWT motif is bound to Exd in only a fraction of the crystal. These two crystals are of lower resolution than the other two crystals (namely, red and black), which do show clear density for the HEWT motif. The failure of the HEWT motif to stably bind to its binding pocket probably contributes to the low resolution seen in these two crystals. It is unclear whether the lack of strong binding seen for these complexes is biologically relevant or just a result of the crystallization conditions.

While there is some correlation between immobilization of the HEWT motif and the NTA in terms of Wilson B factors, this correlation almost disappears when normalizing to

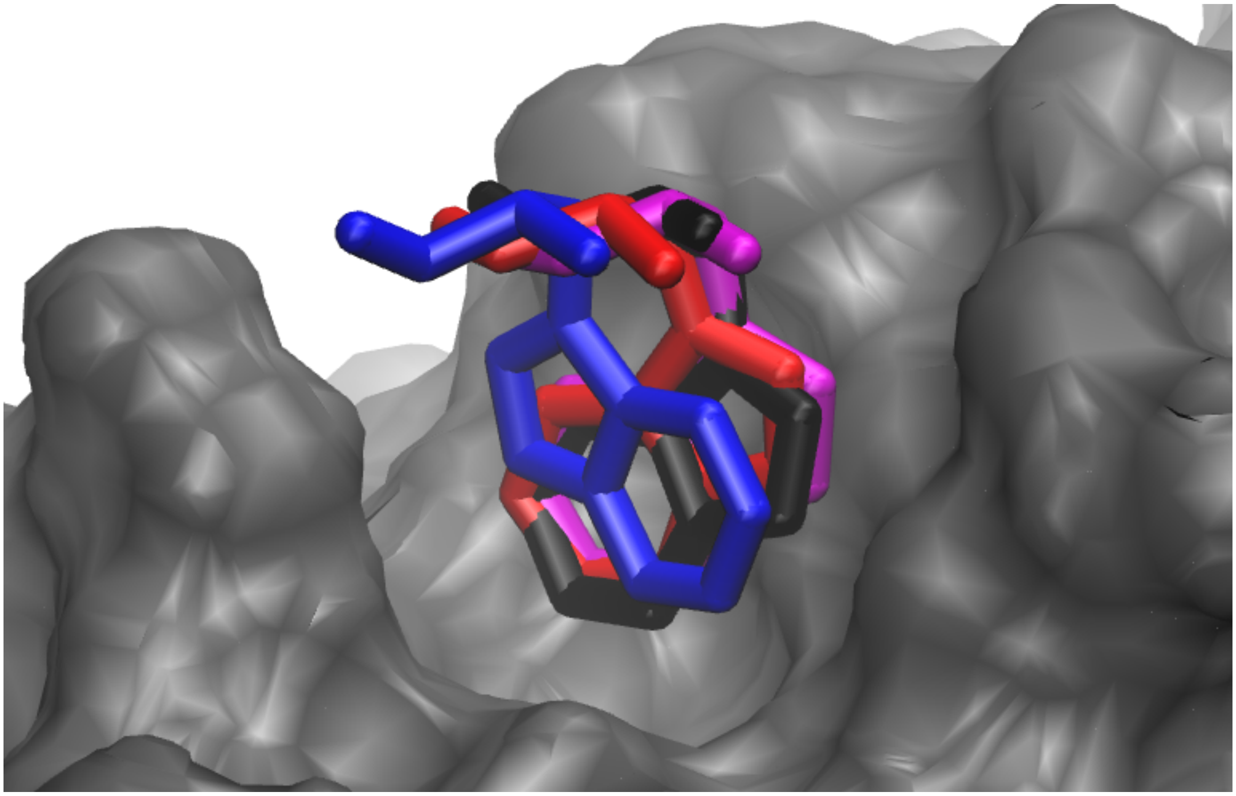


Figure 6.12: **AbdB HEWT tryptophan in its binding pocket on Exd** Superposition of the HEWT tryptophans of all four ternary complexes by RMSD minimization of the Exd backbone. The Exd homeodomain is represented as a gray surface. Only the Exd homeodomain from the black ternary complex is represented for clarity.

overall B factors. The fact that Lys3 is completely absent in the binary complex seen in the black crystal does suggest however that cooperative interaction with Exd plays a role in stabilization of the NTA.

Interestingly, HEWT immobilization does not seem to strongly correlate with binding affinities. The magenta and red sequences are both preferred targets of AbdB, while blue and black are not. The complexes that show clear density for the HEWT motif are red and black however, while blue and magenta crystals show only low occupancy of the motif. As mentioned before this could be merely an artifact of crystallization. In particular, the black crystal, because it features a "red" binding motif in the reverse complementary strand, has an additional binary complex in the asymmetric unit and therefore crystallized in another space group from the other three complexes (P1 vs. C121). This leads to, among other things, different crystal contacts. In particular,the HEWT motif itself is involved in crystal contacts, which could presumably contribute to its immobilization. The magenta complex on the other hand, while having crystallized in C121 just like red and blue, is the only one that crystallized at a higher pH of 9.0, which could potentially affect the binding of the HEWT motif to Exd.

We want to point out that we do not believe that binding of the HEWT motif to Exd is only a result of crystal contacts in the black crystal. The mode of binding that is observed is identical to that reported for many other Hox-Exd/Pbx structures in the literature [49, 108, 117, 129, 133]. We merely believe that crystal contacts further stabilize the bound conformation and possibly falsely inflate its occupancy and deflate its B factors compared to the other structures.

6.8 Preformed vs. induced DNA shape and conformational selection

After addressing questions 5 and 6 from the introduction of this chapter, we analyzed the DNA shape in all five structures to shed light on questions 3 and 4. The program Curves+ was used to analyze DNA shape for all complexes [141] and compared to DNA shape predictions made based on sequence information only with the web based program DNAshape [139]. Figure 6.13 shows a plot of the minor groove widths for all five complexes as measured with Curves+ as well as minor groove widths for the four used DNA sequences predicted by DNAshape. Note that there are only four sequences and thus only four predictions, but that we have an extra binary complex for the black DNA, which we also

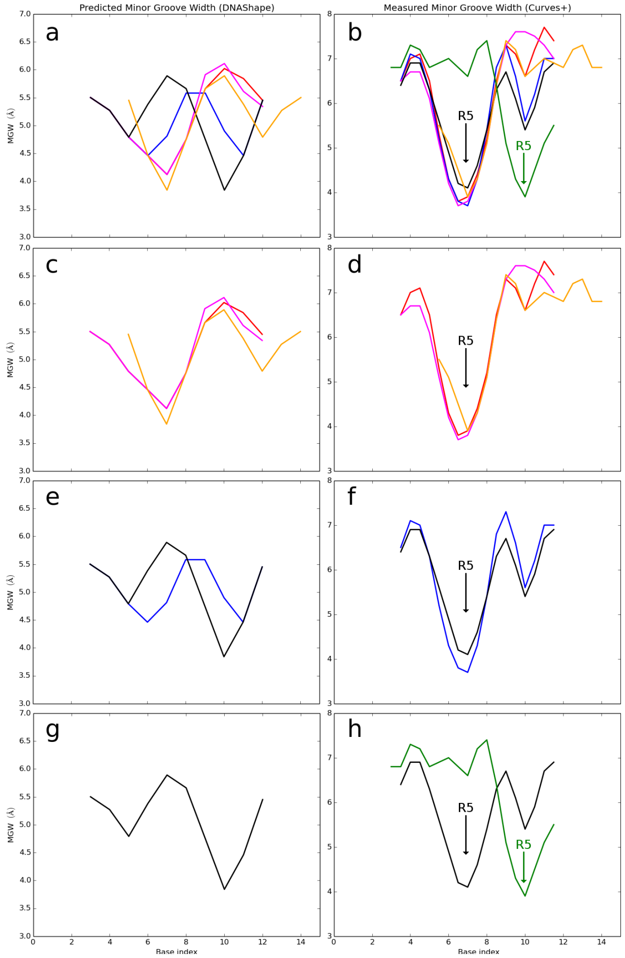


Figure 6.13: **Predicted and measured DNA shapes**
The left column (a,c,e,g) show predicted minor groove widths [139], while the right column b,d,f,h) show minor groove widths measured with Curves+ [140]. The green line in (b) and (h) corresponds to the black DNA in the binary complex. The orange line in (a) and (c) corresponds to the reverse complement of the black DNA when aligned by its red half binding site. The orange line in (b) and (d) corresponds to in the black DNA in the binary complex aligned by its red half binding site in the reverse complement strand. The shape profile is very similar for all measured DNA oligomers (right column) when aligned by Hox binding site [140]. Comparing (c),(d) with (e),(f) shows that the predicted shapes of red, magenta and the red half site in the black DNA matches the measured shape very well. The shape of the black and blue complexes is less well predicted, suggesting a deformation of the DNA upon Hox binding, with a higher energy cost for black and blue than for red and magenta, explaining the higher affinity of AbdB to the latter ones.

analyzed with Curves+. The binary complex in the black crystal is represented twice, once aligned according to its forward DNA strand (green line), such that it aligns with the sequence of the black DNA in the ternary complex, and once according to the red half binding site in the reverse complementary strand (orange), such that it aligns with where AbdB binds in all the ternary complexes.

While DNAshape predicts local minima in the minor groove width at position 1 (A_4T_5 region of the core 12mer) for only the red and magenta oligomers, analysis of the crystal structures shows that all ternary complexes contain a local minimum of around 4Å here. This is where Arg5 inserts into the minor groove for all ternary complexes. This suggests that while a shape propensity can be encoded by the DNA sequence, binding of the complex and possibly insertion of Arg5 in particular, forces this minimum. In stark contrast to this, the black DNA in the binary complex, with AbdB bound in reverse fashion, does not contain this local minimum in the same position but rather has a minimum of the exact same width at position 2 (A_8Y_9 region), where its Arg5 inserts, further supporting the idea of an induced minimum (green line). This result somewhat contrasts the idea described in the literature by which the shape of the DNA is inherent to the DNA, causing different Hox proteins to select different shapes with different preferences [49,118,120,138,142–144]. Instead, at least in the case of AbdB, formation of the ternary complex strongly influences the DNA shape. The local minimum of 4Å perfectly correlates with Arg5 insertion, irrespective of DNA sequence, for both the ternary and binary black complex. If the minimum is “caused” by the insertion of the Arginine “pulling” the phosphate backbone together or by compression because of the “recognition helix” binding on the opposite side of the strand in the major groove, or a combination of both, is currently unclear.

When analyzing the DNA shape at position 2, the measured minor groove widths seem to indeed correlate with the predicted widths [139], with black and blue having narrower minor grooves than red and magenta. This also correlates with the relative binding affinities of AbdB to these sequences [120]. The green line in in the lower right plot represents the black DNA in the binary complex when aligned by sequence to the black DNA in the ternary complex. The minimum at position 2 is clearly reinforced when AbdB binds in reverse fashion, plausibly through the insertion of the Arginine here. The orange line in Figure 6.13 represents the same binary complex but aligned according to its red half binding site in the reverse complementary strand. The similarity in shape with the red sequence immediately stands out when aligned this way, with a strong minimum where Arg5 inserts for both the red and the reverse black DNA and almost no dip in minor groove width at position 2, where no Arginine inserts in the minor groove.

These findings suggest that a strict lock and key mechanism is unlikely, but not that the DNA cannot have a propensity for a narrower minor groove at a certain position along the sequence thus "selecting" for certain Hox proteins ("conformational selection" mechanism). The idea of a conformational selection mechanism is supported by the fact that minor groove widths at position 2 correlate with predicted widths and with the binding preferences of AbdB (seemingly preferring a "wider" minor groove at position 2 but a "narrow" minor groove at position 1, thus preferring red and magenta over black and blue). The fact that the black crystal is the only one where we observe reverse binding, further underlines this idea, as the black sequence is the only one with a somewhat preformed minimum at position 2, possibly priming it for Arg5 insertion and thus reverse binding (*vide infra*).

Even though it seems that the minimum at position 1 is completely absent for the black DNA in the binary complex, the minimum might be somewhat preformed in the other sequences making it easier for AbdB to bind in the canonical fashion with Arg5 inserting into the minimum at position 1. Indeed, the shape predictions show strong minima at this position for the red and the magenta DNAs while predicting smaller or no minima for the blue and black DNAs, which could certainly be part of the explanation for the differences in affinity.

We are now able to respond to our questions 3 and 4 from the introduction of this chapter. It seems like the second minimum (A_8Y_9 region) is somewhat preformed in the different ternary complexes, when no Arginine inserts into the minor groove. The order of widths of the minor groove at this position correlates qualitatively with the predicted widths using DNAshape. The width of the "first minimum" (A_4T_5 region) however seems to be identical for all four ternary complexes and does not correspond the predicted widths. It seems plausible to assume that this is caused by the formation of the ternary complex, either because the Arginine "pulls" the minor groove together due to electrostatic interactions, or because the homeodomains bound to the neighboring major grooves "push" the minor groove together, or a combination of the two effects. In the special case of the binary complex of AbdB bound to the red half site in the reverse complementary strand in the black crystal, this minimum is completely absent, and looks more similar to the predicted shape. Our interpretation is that the black DNA, as predicted by DNAshape, has a propensity to have a minimum at position 2 (A_8Y_9 region) but not position 1 (A_4T_5 region), but formation of the ternary complex "forces" the minimum at position 1. This deformation of the DNA at position 1 costs energy and is only possible because of the extra binding energy provided by ternary complex formation (energy from Exd binding and from cooperative interactions). On the other hand, the propensity of the black DNA to have a narrow minor groove at position 2 but not position 1 allows an AbdB homeodomain to bind to the reverse complement strand without Exd, because less energy is required to deform the DNA and no extra energy is needed from complex formation. In other words, when looking at the black DNA in an upside down fashion, it "looks" exactly like the red DNA in terms of minor groove width, which predisposes AbdB binding to the reverse complement strand, which is why in this specific case we see both binding to the forward and the reverse complement strand. Reverse binding is preferred for the AbdB homeodomain alone because the final DNA shape is somewhat preformed, while forward binding is achieved through deformation of the DNA with energy from cooperative complex formation.

6.9 Exd helix 4

The Exd homeodomain ends at residue 300 of the full length Exd construct. All four of our crystals contain ten extra residues at the C-terminal end. These extra residues have been shown by previous studies to be disordered in solution but to form an α-helix upon binding to the DNA (PDB ID: 1PUF [129], 1LFU [145], 1DU6 [146]). These extra residues are highly conserved among Exd homologs, (Exd, Pbx, Ceh-20) and have been shown to increase affinity to DNA and Hox proteins [147, 148]. All four ternary complexes described herein show clear density after the end of the Exd homeodomain. While appears to have clear helical character, it proved difficult to build a model for this fourth helix. After many rounds of building and refinement, we came to the conclusion that this part of the protein is mostly helical but adopts multiple conformations in the crystal and thus building a single model is difficult. The final models are similar for all four structures and are all partly but not fully alpha helical. B factors for this region are high for all four complexes and all four structures show remaining positive difference density in the proximity of the helix, speaking in favor of additional, unmodelled conformations of this part of the protein.

Most importantly, the model we built for all four complexes is different from that found in the two structures that exist of the vertebrate Exd homolog Pbx that include these additional residues. The helices all lie with their long axes approximately aligned but the exact arrangement of the residues is different. In particular, while the sequence of these 10 residues is very conserved, the main difference is that Exd has an alanine at position +2 (compared to the end of the homeodomain) whereas Pbx has a phenylalanine. In the case of Pbx in the crystal structure 1PUF, the phenylalanine lies across the face of the homeodomain, whereas in the case of Exd the same space is occupied by Gln+3 (compared to homeodomain) which is hydrogen-bonded to a Tyrosine or an Asparagine (or both) depending on the structure, while Ala+2 faces the solvent. This gives the entire helix a shift up. At the same time there is more unmodelled electron density at the C-terminus of the helix in our structures, suggesting that, at least in part of the crystal, the helix is either shifted further down as in the published structures or partly unfolded, thus extending beyond the currently modeled C-terminus. While for 1PUF the electron density is stronger and more clearly defined than for our structures, its B factors are much higher than for the rest of the structure (79 vs. 44 Å^2 for the rest of the homeodomain), speaking in favor of this conformation being not fully occupied as suggested by our own structures. In addition, we examined the C$_\alpha$ secondary shifts (C$_\alpha$ chemical shift minus random coil shift of the particular amino acid type) of the NMR structure of Pbx (PDB ID: 1LFU [145]), which also contains these extra residues. The secondary shifts of this region are about 57% of the average of secondary shifts for the other three helices, further underlining the dynamical character of this helix.

What all structures have in common is that a disordered part of the protein seemingly becomes at least partly structured, with the major conformation being mostly alpha helical and lying across the same face of the homeodomain, at about the same angle compared to helix 3. By doing so, helix 4 has been shown to rigidify the recognition helix 3 [145] and to deepen the binding pocket for the HEWT motif, which could plausibly have functional implications and could possibly fine tune the cooperative interaction through the HEWT motif (see discussion).

6.10 Discussion

Here we have analyzed five structures of AbdB bound to DNA. Four of the analyzed structures were ternary complexes of the AbdB homeodomain cooperatively bound to four different DNA sequences (red, blue, magenta and black) with its cofactor Exd. Coordinate files for all structures will be deposited to the PDB.

At the beginning of this chapter we have asked ourselves the following six questions, which we were able to fully or partly answer herein:

1) Why do Hox paralogs other than Scr that also have an Arginine at position 3 fail to bind these same sequences (in particular fkh, a "blue" sequence)?

2) Why does failure to insert the Arginine result in a failure to bind the sequence?

3) Is the DNA shape predetermined by its sequence and independent of Hox and Exd binding? Is Arginine 3 insertion in the fkh structure a result of the narrow minor groove or the other way around?

4) Do the predicted minor groove shapes hold up when examined with X-ray crystallography and is there a clear correlation between actual minor groove width and Hox affinity?

5) Is a local minimum in the DNA sufficient for insertion of residue 3 into the minor groove? That is, will posterior Hox proteins also insert residue 3 into the local minimum of a sequence that is not a preferred binding site?

6) If the shape of the DNA dictates the conformation of the NTA for anterior Hox proteins like Scr, does this apply to other Hox proteins as well, in particular posterior ones? In other words, are all Hox proteins equipped to recognize DNA shape or only some?

A discussion of the findings with regard to these questions, as well as additional findings follows below.

6.10.1 Why do not all Hox proteins containing an Arginine at position 3 bind to sequences with a second minor groove width minimum

This question was addressed both in chapter 5 and 6. The answer seems to be that the residues at position 4 and 6 of the N-terminal arm contribute to the insertion of residue 3. Threonine 6 in the case of Scr (and presumably other *Drosophila* Hox proteins that have a Threonine at position 6, namely lab, pb, and Dfd) appears to form a hydrogen bond with the phosphate backbone of the DNA, the energy of which may contribute to the compensation of the energetic cost of the "twisting" of the NTA as well as the entropic cost of inserting Arg3 into the minor groove. This is supported both by our analysis of MD simulations of Scr and Scr mutants on fkh (chapter 5), as well as comparison of our crystal structures to previously published crystal structures [49]. This conclusion also supports *in vitro* and *in vivo* studies of Scr mutants of position 4 and 6 reported in the literature, which have reduced capability to select DNA sequences with this second minor groove minimum [118]. This would suggest that the four Hox proteins which have a threonine at position 6 (Scr, Dfd, pb and lab) would show a preference for DNA sequences with a second minimum, like green and blue sequences. This seems to be the case as shown in the original SELEX-Seq results [120].

The influence of position 4 is a little bit more ambiguous but our MD simulations suggest that replacing the glutamine of Scr with a glycine (as seen in Ubx and Antp) has a somewhat weaker but additive negative effect (in case of the double mutant) on Arg3 insertion. This is possibly due to the entropic cost of immobilizing Arg3 when directly preceded by a Glycine. The absence of electron density for Gly4 in a crystal structure of Ubx and Exd bound to DNA seems to support the idea that Gly4 is highly flexible and the insertion of Arg3 into the minor groove would thus have a higher entropic cost than for Scr, where backbone electron density for Gln4 was seen in the structures with both the specific and the consensus binding sites [49, 108].

In the case of AbdB, which is the only Drosophila Hox paralog with a lysine at position 4, we see an interaction of the lysine with the DNA backbone, which plausibly contributes both to the failure to insert residue 3 into the minor groove as well as to the posterior prevalence phenomenon [129–131], by which more posterior Hox proteins can outcompete more anterior ones (another contributor likely being its shortened linker region and thus smaller entropic cost of complex formation).

6.10.2 Why does failure to insert Arginine 3 lead to decreased affinity

We were not able to fully address this question but evidence from this and the previous chapter (chapter 5), as well as comparisons between different published Hox-DNA complexes suggest that the energy of the interaction of Arg3 with the minor groove compensates for other energetic costs, possibly entropic costs of deforming and immobilizing the DNA (*vide infra*). For example, Joshi et al. [49] suggested that the consensus sequence fkhCON (related to our red sequence) not only lacks the second minimum and has a more pronounced first minimum, but is also much less flexible in solution than the specific Scr target fkh. A plausible explanation would be that the energetic cost of binding to fkhCON, which is more rigid in solution, is small enough for all Hox paralogs to successfully bind to it, regarless of the identify of the residues in the NTA. The specific fkh sequence however (related to our blue sequence), seems to be more flexible in solution and the entropic cost of binding and thus reducing its flexibility, can only be balanced by the energetically favorable interaction of Arg3 with the DNA minor groove, if the conformation of the NTA allows for such insertion, which seems to be determined by the identity of other residues in the NTA, in particular positions 4 and 6 (discussed above and below). The energetic cost of DNA deformation will be discussed more below.

Alternatively (or additionally), the fact that Arg3 inserts into the minor groove could put certain restraints on the linker region N-terminal to it, which could possibly modulate its affinity to Exd and thus complex formation.

6.10.3 Is DNA shape independent of Hox and Exd binding

We were able to answer this question unequivocally by analyzing our five structures. A qualitative correlation is observed between predicted and observed DNA shape for the minor groove in the A_8Y_9 region of all ternary complexes. The minor groove width at this position also correlates with binding affinity of AbdB to the respective sequences. However, the minor groove width in the A_4T_5 region seems to be determined by formation of the complex. The minor groove width here was almost exactly 4Å regardless of the DNA sequence for all ternary complexes (where Arg5 inserts in that region). This minimum is completely absent for the binary complex that is missing Exd and does not insert its Arginine in the same position. Instead the binary complex has a minimum of also exactly 4Å in the A_8Y_9 region, where it inserts its Arg5. These findings strongly support the idea that at least in the case of AbdB, the binding of the homeodomain determines the shape of the minor groove. However, as mentioned above, this only disproves a strict lock and key binding mechanism for shape recognition but does preclude a conformational selection mechanism. The fact that only the black sequence, which has a preformed minimum in the A_8Y_9 region, crystallized with the AbdB homeodomain bound in reverse fashion and in the absence of Exd, suggests that this preformed minimum allows AbdB binding to this half site to compete with ternary complex formation on the reverse complementary strand, the preformed shape offsetting the missing energy from cooperative binding with Exd.

6.10.4 Can predicted DNA shapes be confirmed by crystal structures and is there a correlation with affinity

As mentioned in the previous section, the so-called first minor groove minimum in the A_4T_5 region seems to be fully determined by Hox binding. The shape of the DNA in the A_8Y_9 region however seems to qualitatively reproduce the predicted DNA shapes based on sequence alone. This suggests that while these local minima in the minor groove are not solely determined by the sequence, different sequences can have different shape propensities in solution which are then preferentially bound by the respective Hox proteins. As mentioned before this corresponds to a mixed induced fit/conformational selection mechanism, which can partly explain the affinity differences of the different *Drosophila* Hox paralogs.

6.10.5 Do posterior Hox proteins insert residue three into a preexisting minor groove minimum

In none of the five structures described here do we see density in the minor groove in the A_8Y_9 region, which would suggest insertion of Lys3 into the minor groove. The blue and black sequences, which have preformed minima in this position, show the N-terminal arm in exactly the same conformation (with no insertion into the local minimum) as in the red and magenta complexes which lack this preformed minimum. Additionally, the observed conformation in all our structures is identical with the conformation of a mammalian AbdB homolog described in the literature [129].

As suggested by the comparison with the crystal structures described in this chapter with other structures described in the literature [49] [129], a well as by the MD simulations in chapter 5, it seems that in particular the residues in positions 4 and 6 of the N-terminal arm are responsible for this. Posterior Hox proteins lack the threonine at position 6 which is seen for anterior Hox proteins and in crystal structures interacts with the DNA backbone, which allows the N-terminal arm to adopt a conformation primed for the insertion of Arg3 into the minor groove. The common NTA conformation is not only shared by all AbdB structures described herein and its homolog HoxA9 [129], but also for another posterior *Drosophila* Hox protein (Ubx), which lacks Threonine 6 (PDB ID 1B8I [108]). In addition, in the case of AbdB, the Lysine at position 4 interacts with the DNA backbone and possibly contributes to the failure of residue 3 to insert into the minor groove.

6.10.6 Are all Hox proteins similarly shape sensitive

The data presented herein suggest that not all Hox proteins are similarly shape sensitive. All AbdB structures described herein, be it with or without Exd, show their NTA in the exact same conformation. This conformation is identical with the ones seen in the literature for the AbdB homolog HoxA9 [129] and the Drosophila Hox protein Ubx [108]. This conformation is identical regardless of the DNA target. Two of the DNAs used herein are predicted to have a second minor groove width minimum in the A_8Y_9 region, which is qualitatively reproduced in the crystal structures. While the order of the minor groove width is qualitatively correlated with the binding affinities of AbdB to these four sequences, we failed to find a structural explanation that would account for the different affinities, at least when considering the A_8Y_9 region.

However, when we look at the A_4T_5 region, DNA shape prediction indicated the red and magenta sequences, which have high affinities to AbdB [120], to have preformed minima, while the blue and black sequences do not, or less so. The preference of AbdB to the red and magenta sequences could thus be explained by a conformational selection mechanism, which leads to a preference for oligomers with a preformed minimum in the A_4T_5 region. The fact that we saw an additional binary complex in the black crystal, with AbdB recognizing a red half site in the absence of Exd, supports this idea, especially considering that Hox binding in the absence of Exd is reportedly much weaker [112] [116], meaning that the energetic gain from using a preformed minimum must be compensating for the energy lost from cooperative binding with Exd.

An additional explanation could be that not only the shape of the DNA matters but also its flexibility in solution. The flexibility of the DNA will be restrained upon complex formation, resulting in an entropic cost that has to be compensated by energetically favorable interactions between the components of the complex. For instance, in the case of fkh, which is related to the blue sequence described herein, not only did Monte Carlo simulations suggest a more narrow minor groove in the A_8Y_9 region, but also a greater conformational flexibility [49]. This could mean that in addition to the preformed minimum in the A_4T_5 region having a favorable energetic contribution, a favorable entropic contribution from the A_8Y_9 region could contribute to the preference of red and magenta sequences over black and blue ones.

This idea is supported by mutational studies [49]. Mutating Arg3 of Scr to an Alanine, reduces the affinity of Scr to fkh but not to fkhCON. This could mean that binding to fkhCON is inherently energetically more favorable and binding to fkh can only be achieved through energetic compensation of Arg3 inserting into the minor groove. Following the same logic, any Hox protein should in general be able to bind a red sequence more easily than a blue sequence, both because of the stronger preformed first minimum described herein, and the higher rigidity of the red DNA in solution. A general preference for red sequences among Drosophila Hox proteins seems to be supported by the data reported in the literature [120].

In the case of sequences that are more like fkh and thus blue sequences, which shows higher conformational flexibility in solution but at the same time has a tendency to have a narrower minor groove in the A_8Y_9 region, additional energy from Arginine 3 insertion but also contribution from other NTA residues is necessary to compensate for the higher entropic cost, in particular Threonine 6 in the case of the more anterior Hox proteins(Scr, Dfd, pb and lab), which can contribute by interacting with the DNA backbone, and Glutamine 4, which is probably simply reducing the entropic cost of Arg3 insertion as compared to a Glycine at the same position. Scr and Dfd, which are the only two Hox proteins with very high affinity to blue sequences, are also the only two Hox proteins in the fly with both a Threonine at position 6 and a Glutamine at position 4. This classification according to residues 4 and 6, corresponds exactly the the three major classes of Hox proteins as classified by their preference to either the red, the blue or the green core motif [120], class 1 consisting of lab and pb (Threonine at position 6, but no Glutamine at position 6; pb actually has a Leucine at position 6, which if we use the entropic argument for position 4, explains its remaining affinity for blue sequences), class 2 of Scr and Dfd (Threonine at position 6 and Glutamine at position 4) and class 3 of all the posterior Hox proteins, which all lack the Threonine at position 6 (Ubx, Antp, AbdA and AbdB). The importance of these two residues is further supported by mutational studies. An Scr mutant with the mutations T6Q and Q4G lost its preference for blue sequences and instead behaved much like the posterior Hox protein Antp in SELEX-Seq experiments [118] and biochemical assays shows that the Q4G mutation leads to a ∼ six-fold reduction in affinity of Scr to fkh [49].

This could be a general structural explanation for the findings presented herein and in the literature, that explains the binding preferences seen for the different Hox paralogs as well as the relationship with the identities of the amino acids in the N-terminal arm. It explains the three classes of Hox proteins in the fly described in [120] and also explains why Dfd and Scr (class 2) that have both the entropically favorable Glutamine in position 4 and the enthalpically favorably Threonine at position 6 are the most promiscuous binders. Posterior Hox proteins probably still outcompete anterior ones for general target sites like the red sites, because of their shorter linker regions (lower entropic cost of binding cooperatively with Exd) as well as additional mechanism not described herein, like additional interaction motifs [133, 149]. To confirm this mechanism, one would have to confirm the differences in conformational entropy of the different types of DNA by Monte Carlo and/or MD simulations and experimentally by NMR spectroscopy in the absence or the presence of the homeodomains of Hox proteins and Exd.

6.10.7 Possible functions of Exd helix 4

The ten extra residues C-terminal to the Exd homeodomain were shown to adopt a helical conformation for all four ternary complexes, similarly as described in the literature [129, 146]. The exact conformation of this helix in the structures described here is different from the two examples described in the literature however, which were of the mammalian Exd homolog Pbx. All our structures as well as the structures in the literature suggest that these residues are only partly helical and possibly adopt multiple conformations in the crystal, as suggested by relatively high B factors for all structures, unmodeled extra density and not fully alpha helical secondary chemical shifts.

The helical structure forms upon binding to the DNA and thereby rigidify the DNA recognition helix 3 and deepening the binding pocket for the HEWT motif [129,145]. In this way helix 4 could modulate the interaction with the DNA and the Hox protein [147, 148]. These residues are very conserved among Exd and its homologs but do not seem to adapt the exact same conformation upon binding to the DNA for the different complexes described here and in the literature [129, 145]. Additionally in all the described structures they seem to retain some flexibility and possibly occupy multiple conformations. These findings make this region an interesting candidate for regulatory roles. It could potentially relay sequence information through the recognition helix, which it is attached to, to the HEWT/YPWM motif or to other proteins that are involved in the transcription process *in vivo*.

Interestingly, in the two Scr structures described in the literature [49], the Hox linker extends to the region where Exd helix 4 would lie according to our and other structures. This suggests that if helix 4 were present in these complexes it would force the Hox linker to adopt different conformation and very possibly interact with it. Such an interaction could modulate the cooperativity between Hox and Exd and could potentially even lead to the formation of additional structures. Depending on the identity of the linker in the different Hox paralogs and the target DNA the complex is bound to, different structures could form that can be recognized and further regulated by other cellular elements. Such mechanism could potentially tune expression of the gene in one way or another. Hox proteins have been shown to both activate and repress the genes that they bind to [150, 151]. The type of additional structures formed could determine the regulatory output of the complex formation (*vide infra*).

One particularly interesting interaction can be seen in the black ternary complex only. The histidine that is part of the HEWT motif seems to directly interact with helix 4 through a hydrogen bond (possibly water mediated). This interaction is not seen in any of the other crystals described herein or in the literature so it is unclear whether it plays a functional role *in vivo*, in particular because the same histidine is involved in crystal contacts. But the fact that the HEWT and in particular the tryptophan occupy the binding pocket on Exd in almost identical fashion in this as compared to all previously published structures and helix 4 lies close enough to even form this direct interaction makes it plausible that this interaction is not simply an artifact and might be seen at least in some sub ensemble of complexes in the solution. This interaction could for example represent an intermediate state in cooperative complex formation. This would have to be investigate further, for example by mutational experiments and/or solution state NMR.

While the position of the tryptophan in the binding pocket of Exd seems to be conserved across all structures, the conformation of the surrounding residues is slightly different in the different structures also when compared to previously published structures [49,108,117, 129, 133]. It has been suggested in the literature that the linker conformation might affect the regulatory output of Hox binding to its target site. For example while there are certain binding sites that can be bound by different Hox paralogs, their effect on gene expression might be different (both positive and negative [150, 151]). The conformation of the linker, and thus the regulatory output, might not only be influenced by the type of the linker but also be influences by how the hexapeptide binds to Exd, which in turn could be modulated by both the conformation of the NTA and interactions with other parts of the protein, for example Exd helix 4. The idea of helix 4 being able to modulate regulatory output is interesting also because it is connected to the DNA recognition helix (helix 3) and is able to adopt different conformations in the different crystal structures, while being largely disordered in solution [146]. It is conceivable that helix 4 could be a way for Exd to relay information about the DNA from the recognition helix to the hexapeptide and thus the Hox linker region.

6.10.8 Possible competition between overlapping binding sites

The cocrystallization of a binary and a ternary complex in the asymmetric unit of the black crystal not only provided us with the possibility to directly compare AbdB binding with and without its cofactor Exd, but also offers some room for speculation about the possibility of overlapping binding sites competing for different Hox proteins. Binding sites could be overlapping on the same strand of DNA or on reverse complement strands as seen for the black DNA in the present study. An analysis of the SELEX-Seq data that was previously published [120], which was kindly provided by the laboratory of Dr. Richard Mann, confirmed that up to 15% of all sequences selected for some Hox proteins contained multiple Hox-Exd binding sites. The black DNA is particularly likely to contain a red sequence in its reverse complement strand, either a red Hox only half-site or a full red

Hox-Exd binding site, depending on the nucleotides flanking the core 8mer. The green DNA in the other hand seems particularly likely to contain a second Hox-Exd binding site on the same DNA strand, because its core 8mer is simply a tandem repeat of the preferred Exd binding site TGAT.

This begs the question if overlapping binding sites play a role *in vivo*, where different Hox proteins or Hox-Exd complexes might be competing for the same space on the DNA. We have seen herein that AbdB binding to the red half-site on the reverse complement strand of the black DNA occupies much of the space that would otherwise be bound by Exd thus acting as a competitive inhibitor for Hox-Exd complex formation on the forward strand. In particular, the black sequence used in this study, has recently been reported to appear in a cluster of low affinity Hox binding sites in the *shavenbaby* enhancer [152]. Since the reported binding sites are low affinity binding sites and we have seen in the present study that is possible for a high affinity Hox monomer binding site to compete with a low affinity Hox-Exd binding site that overlaps with it, it seems plausible to assume that such competition would be seen *in vivo*. This could mean that depending on the stage of development and the segment of the embryo, the enhancer could be active or inactive depending on how well the Hox protein present in a particular segment at this stage of development competes with its Hox-Exd counterpart for the same space on the DNA. How well the Hox protein competes with the Hox-Exd complex might depend on the identity of the Hox protein as well as the identity (sequence and shape) of the competing binding sites and the expression levels of Exd.

One could perform gel shift assays with the *shavenbaby* enhancer with the different Drosophila Hox proteins at different ratios of Hox to Exd to evaluate this hypothesis. One would evaluate the ability of the binary complex to form even in the presence of Exd. The next step could be a similar experiment where a second Hox protein is added to determine if the ratio can be shifted towards binary complex formation. The results would indicate whether it is possible for any of the Hox proteins to compete with their own or other Hox-Exd counterparts for overlapping binding sites *in vitro*. If Hox proteins can be identified for which Hox binding successfully competes with binding of Hox-Exd, one could try to see if ectopic expression of those Hox proteins can competitively inhibit *shavenbaby* expression in segments where it is usually active. For example, *shavenbaby* expression is activated in segment A1 by the Hox paralog Ubx. Ectopic expression of other Hox paralogs such as Scr or AbdB in this segment may reduce the expression of shavenbaby by competitive inhibition for overlapping binding sites in the *shavenbaby* enhancer.

Chapter 7

Conclusions and future directions

7.1 Reproducibility of molecular dynamics derived order parameters

This section was published, in part:

Zeiske T, Stafford KA, Friesner RA, Palmer AG (2013) Starting-structure dependence of nanosecond timescale intersubstate transitions and reproducibility of MD-derived order parameters. Proteins: Structure, Function, and Bioinformatics 81: 499 - 509. [34] Reprinted with permission from John Wiley and Sons.

In chapter 3, we used MD simulations of GB3, a common model system for protein dynamics, for a detailed investigation of discrepancies between different sets of simulations and between simulations and NMR spin relaxation experiments. We compared the square of the generalized order parameter of the backbone NH bond vector derived from MD simulations and NMR spin relaxation measurements. Major discrepancies between different sets of simulations are due mostly to flexible regions of the protein undergoing nanosecond timescale motions corresponding to transitions between subensembles in conformational space. The four glycines of GB3, all situated at the end of secondary structures at the loop hinges, consistently are outliers in the different simulations.

Nanosecond timescale transitions involve movements of flexible regions of the protein, such as the loops and termini, and are often coupled to the breaking or forming of hydrogen bonds. The autocorrelation function of the involved NH bond vector does not converge for simulation trajectories that are not much longer than the timescale of these transitions. Thus, the effect on the average order parameters is significant for numbers of simulations that can be run on current commodity computer clusters, making sampling a predominant determinant for the agreement of different simulations to each other and to order parameters obtained by NMR spin relaxation experiments. Improved agreement with experiment for order parameters averaged over the 1.2-microsecond trajectory supports this conclusion. However, not all discrepancies observed between MD simulations are resolved by increased sampling. The example of the hydroxyl group of Tyr3 illustrates the strong dependence on the starting structure even for very long simulations. It also highlights the need of care in preparing the structures for simulation, especially when adding hydrogen atoms to the X-ray crystal structures, which usually lack hydrogen atoms.

The differences between starting structures do not have to be large or obvious to have a noticeable influence on the dynamical behavior of the protein. Earlier studies focused on starting structures with different backbone conformations or starting structures derived from different crystal structures [60, 61]. Here we showed that seemingly small conformational differences in sidechains resulting from different setup protocols applied to the same crystal structure can influence the backbone order parameters for sites distant in sequence or space and therefore seemingly uncorrelated at first glance. Furthermore, we identified specific molecular interactions responsible for the altered conformational dynamics for different starting structures of GB3. In summary this study emphasizes the importance of both increased sampling and good choices of starting structures in MD simulations. Eliminating nanosecond timescale motions when averaging order parameters over all simulations increases agreement between simulations and experiment. However, force field or sampling limitations might not be the only issues in accurately characterizing nanosecond or slower motions, because NMR spin relaxation techniques are largely insensitive to motions in this time regime. Thus, full understanding of the processes that can be captured by NMR measurements are necessary when judging the accuracy of MD simulations. Identifying and understanding the discrepancies and aberrances between simulations and NMR experiments can provide insights that help develop better force fields or NMR experiments and improve their interpretation.

7.2 Thermostability of enzymes from molecular dynamics simulations

In chapter 4, we examined Molecular Dynamics (MD) simulations for a number of orthologs from the RNase H family of enzymes. The premise of the chapter was to determine if temperature changes in simulations enable us to calculate thermodynamic parameters that reflect parameters obtained by experimentation. When simulating RNase H constructs from a range of organisms, including psychrotrophic, mesophilic and thermophilic organisms, we were surprised to find a very linear relationship of experimentally determined melting temperatures T$_m$ with the dimensionless parameter $\Lambda = dln(1-S)/dlnT$, where S is the order parameter of the backbone NH bond vector. One implication of this result is that the melting temperature is mainly determined by the temperature dependence of the backbone but not the side chains. For the heat capacity C$_p$, however, no linear relationship to the backbone Λ values could be determined, which may imply that the side chains play a larger role than for the melting temperature.

Merely one homologous family of proteins was studied here. While the results are promising, other protein families have to be examined to confirm that the observations made here are generalizable. If linear relationships are observed for other systems, MD simulations could become a powerful tool for protein design. Because no full folding and unfolding events need to be simulated, many point mutants of a certain protein domain could be screened relatively quickly for increased thermostability. The best candidates could then be expressed and tested experimentally. This approach would also complement well with high throughput protein engineering methods described elsewhere [83].

7.3 Hox specificity

In chapters 5 and 6 of this dissertation, we studied the DNA-protein complex formation of the Hox transcription factor family. Chapter 5 used NMR spectroscopy and MD simulations to study the *Drosophila* Hox paralog Scr together with its cofactors HM and Exd, while chapter 6 used X-ray crystallography to study the *Drosophila* Hox paralog AbdB together with its cofactor Exd. Both chapters contribute to our understanding of the specificity of Hox-DNA interactions in general, and of Hox-DNA interactions in particular.

In chapter 5 we have developed expression and purification protocols for several components of the *Drosophila* "Hoxasome" complex [150], including the homeodomains of Scr, AbdB, Exd as well as the cofactor dimer HM/Exd. We have also laid the groundwork for future studies of Hoxasome complexes, by identifying complex formation conditions and establishing methodology to asses complex formation. In particular, we recorded 1D NMR spectra of unlabeled DNA oligomers specific to the *Drosophila* Hox protein Scr. Peaks in the imino proton region of the DNA have been assigned using NOESY experiments, and line widths and T_2 relaxation times measured using Hahn echo experiments to assess tumbling rates of the DNA in solution. Hahn echo experiments were also performed on DNA in the presence of Hox protein or cofactors, and binding could be confirmed by comparing transverse relaxation times. Importantly, salt titrations could confirm a ternary complex of Scr, Exd and DNA to be in slow exchange speaking in favor of a cooperative interaction, when compared to binding of just one of the two protein components to DNA, a process which seems to be in the fast exchange regime.

Preliminary 2D experiments on Scr suggest conformational changes for both the protein backbone (HSQC experiments) and side chains (methyl TROSY experiments) upon binding of Scr to the DNA. More importantly, analysis of shift perturbations of the Scr methyl peaks upon addition of Exd, suggest that interaction with the cofactor leads to additional changes in the Hox protein, presumably including the linker region. Peaks that we have tentatively assigned to the Scr linker region are perturbed upon addition of Exd to a prebound Scr-DNA complex. Further experiments are needed to unequivocally assign Hox sidechain and backbone peaks. Once assigned, relaxation studies on the complex and in particular the NTA and linker region could yield useful information about conformational changes and disorder-order transitions upon binding of DNA and subsequent binding of the cofactor.

At the end of chapter 5, we have conducted MD simulations with the aim to understand the role played by particular NTA residues in Hox specificity. The simulations were performed on wildtype and mutant constructs of Scr bound to its specific target sequence fkh and a the consensus variant fkhCON, based on crystal structures described in the literature [49]. Mutations of both residue 4 and residue 6 in the NTA lead to a decreased population of the canonical conformation of Arg3 inserted in the the minor groove as observed in the crystal structure. A strong correlation is seen in particular for a hydrogen bond of Thr6 with a DNA phosphate. Breaking this bond strongly correlates with Arg3 leaving its canonical inserted state, suggesting the possibility that the NTA is in a somewhat strained conformation when inserting Arg3 into the minor groove. The favorable energy from the hydrogen bond very possibly offsets the cost of this unfavorable conformation. Importance of both residues 4 and 6 for the specificity of Scr has been confirmed by mutational studies in the literature [49, 118]. A caveat to these observations is that MD simulations were performed with a construct based on the published crystal structures, which do not include the linker region of Scr [49]. While we do believe that a qualitative analysis of the Arg3 populations is meaningful, a quantitative analysis might not be. Simulations of Scr including the linker region could be performed to confirm the findings presented here, but could also lead to an additional bias in the simulations, when the linker is built in a non native confirmation, because the structure of the linker region is not known and would have to be modeled without a template. If the linker is included in simulations, one would potentially need to also add the cofactor Exd to the simulations to conformationally restrain the movements of the linker region. Adding both linker and Exd adds computational costs to the simulations.

In chapter 6, we used X-ray crystallography to study the DNA binding preferences of the *Drosophila* Hox paralog AbdB, the most posterior of the fly Hox proteins. The structures of four ternary AbdB-Exd-DNA complexes and one AbdB-DNA binary complex were solved for four different DNA oligomers: red, blue, magenta and black sequences, named after their core octamer as described in the literature [120]. As a posterior Hox protein, AbdB prefers red and magenta binding sequences over black and blue ones. The asymmetric unit of the crystal made using the black DNA oligomer, contained a ternary complex of AbdB, Exd and DNA and an additional binary complex of AbdB and DNA without the cofactor Exd. The reverse complement strand of the black DNA contains an additonal red half binding site but is missing the Exd half binding site. The fkh and fkhCON sequences described in chapter 5 and the literature [49] are similar to the blue and red sequences described here.

When analyzing the minor groove width (MGW) of all five complexes described herein (Figure 6.13), it appears as though the width profile is almost identical for all ternary complexes. A strong minimum of MGW is observed in all ternary complexes where Arg5 inserts its sidechain into the DNA minor groove. A minimum of the same width is observed for the extra binary complex, in which AbdB binds the red half site in the reverse complement strand of the black DNA. When comparing the measured DNA shapes to the shapes that were predicted by the DNAShape server [139], based on sequence alone, it becomes apparent that the predictions match the measured DNA shapes very well for the red and magenta sequences, as well as the reverse complement black DNA, when aligned by its red half binding site. In other words, if the prediction server accurately describes the DNA shape in solution, prior to Hox binding, the shape of the red and magenta sequences (and the red half site in the black sequence) are preformed for Hox binding. This suggests a lower energetic cost for AbdB to bind to the red and magenta sequences than to the black and blue sequences. This confirms the observed preference for AbdB to red and magenta sequences [120]. Such a mechanism would fall under the "conformational selection" binding mechanism, and has been observed for drug binding to the minor groove of DNA oligomers [153].

The case of the extra binary complex of AbdB bound to the black DNA is particularly interesting. Comparing the MGW of the black DNA in the binary and ternary complex shows strong differences between the two when aligned by sequence of the forward strand. When comparing both MGW profiles to the predicted MGW, it seems as though the shape of the DNA in the binary complex is much closer to the predicted shape than the shape of the DNA in the ternary complex. This would mean that the energetic cost of binding to the red half site in the reverse complement strand is lower than that of binding to the black Hox half site in the forward strand. The only reason AbdB is also seen to bind to the forward strand is because of the extra energy from cooperative complex formation with its cofactor Exd. In the case of the red and magenta DNAs the preferred shape profile is preformed in the "forward direction", which also includes an Exd binding site. Cooperative interaction with Exd only increases the affinity to the already preformed preferred shape, explaining the high affinity to those two sequences, whereas for the black DNA cooperative interaction for the forward site is competing with the preferred shape on the reverse complement site. This competition between two binding sites could be of biological relevance but has not been studied further herein. An analysis of the SELEX-Seq data, kindly provided by the laboratory of Dr. Richard Mann, shows overlapping binding sites to be of non negligeable frequency. Affinities will have to be measured for such binding sites for both Hox only and Hox-Exd binding to further assess this question. Gel shifts and *in vivo* competition experiments of different Hox proteins for overlapping binding sites could then be performed to study a possible biological role for binding site competition. Of particular interest for such studies would be the shavenbaby enhancer which contains a sequence identical to our black sequence and thus contains overlapping Hox and Hox-Exd binding sites [152].

We also analyzed the conformation of the NTA in the different structures solved here and compared it to the conformations seen in other structures reported in the literature. All five (four ternary and one binary) complexes show the NTA in a very similar conformation. Arg5 inserts into the minor groove, while Lys3 extends towards helix 3 of AbdB. This conformation is also observed for the previously published structure of the mammalian AbdB homolog HoxA9 bound to DNA together with the mammalian Exd homolog Pbx.

Comparisons with the two crystal structures of Scr bound to its specific target fkh or the consensus target fkhCON however show very different conformations. As stated above, Scr is the only *Drosophila* Hox paralog able to bind fkh *in vivo*. This has been linked to the insertion of Arg3 into a local minor groove width minimum. A set of very different dihedral angles in the NTA are necessary to accomodate this insertion. As stated in the context of the MD simulations performed here on Scr, as well as described in the literature [49, 118], residues 4 and 6 play an important role in Hox specificity. The amino acid identity at these two positions also corresponds well to the three specificity classes described before [120].

Scr and Dfd for example are the only two Hox paralogs with a Glutamine at position 4 and a Threonine at position 6 and show very similar affinity fingerprints in the SELEX-Seq experiments. As described in the context of the MD simulations, the crystal structures of both Scr on fkh and on fkhCON show Thr6 in hydrogen bonding distance to the DNA phosphate backbone. Both structures have a certain “kink” in the NTA backbone at this position, very probably related to this interaction. This leads to differences in NTA backbone dihedrals, leading to Arg3 being in a position that is more poised to insert its side chain into the minor groove in the case of fkh (blue-like sequence), whereas it remains disordered in the crystal structure of Scr with fkhCON (red-like). While the minimum in the minor groove width, and the associated electrostatic potential [49,138,142], is probably necessary for the insertion itself, the hydrogen bond of Thr6 with the phosphate backbone and the ensuing “twist” on the NTA seem necessary for the “poised” position of Arg3. While Arg3 is the residue used for shape readout [49, 118], only the interaction of Thr6 with the DNA backbone “enables” the DNA shape sensitivity of the NTA, making Hox paralogs with a Threonine at position 6 more shape sensitive than the ones that do not. Namely, this would make the anterior Hox proteins more shape sensitive than the posterior ones, at least for the “second” minimum in the minor groove, which seems to indeed be selected by anterior Hox proteins [49, 118].

Why are posterior Hox proteins unable to bind sequences that have a propensity for two minima? One reason could be that the first minimum is not as deep as for the more preferred red and magenta sequences, which seems to the the case for the sequences studied herein. An additional factor could be the higher conformational flexibility seen for the blue-like fkh sequence in Monte Carlo simulations [49]. This would add an entropic cost to the binding of such a sequence by any Hox protein, in addition of any energy needed to “deform” the DNA. This entropic cost can be overcome by Scr because of the interactions of Thr6 with the backbone of the DNA and the insertion of Arg3 into the minor groove. For posterior

Hox proteins, where those two interactions cannot form, the entropic cost of binding a very flexible DNA sequence might be too high.

To separate entropic and enthalpic effects of DNA binding, as well as to confirm the different shape propensities of the different oligomers before protein binding, NMR experiments and constrained MD simulations could be performed. A combination of NOE, RDC and relaxation measurements on the DNA before and after protein binding could be performed in the future, entropic and enthalpic components quantified and a general quantitative theoretical model for Hox specificity established.

Bibliography

[1] Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H, et al. (1958) A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. Nature 181: 662–666.

[2] Vojtechovsk J, Chu K, Berendzen J, Sweet RM, Schlichting I (1999) Crystal structures of myoglobin-ligand complexes at near-atomic resolution. Biophysical Journal 77: 2153–2174.

[3] Wikipedia (2011). Bragg's law — Wikipedia, the free encyclopedia. URL `https://en.wikipedia.org/wiki/Bragg%27s_law`. [Online; accessed 10-November-2015].

[4] Wthrich K, Wider G, Wagner G, Braun W (1982) Sequential resonance assignments as a basis for determination of spatial protein structures by high resolution proton nuclear magnetic resonance. Journal of Molecular Biology 155: 311–319.

[5] Palmer AG (1993) Dynamic properties of proteins from NMR spectroscopy. Current Opinion in Biotechnology 4: 385–391.

[6] Palmer AG (1997) Probing molecular motion by NMR. Current Opinion in Structural Biology 7: 732–737.

[7] Palmer AG (2001) Nmr probes of molecular dynamics: overview and comparison with other techniques. Annual Review of Biophysics and Biomolecular Structure 30: 129–155.

[8] Palmer AG (2004) NMR characterization of the dynamics of biomacromolecules. Chemical Reviews 104: 3623–3640.

[9] McCammon JA, Gelin BR, Karplus M (1977) Dynamics of folded proteins. Nature 267: 585–590.

[10] Lindorff-Larsen K, Piana S, Dror RO, Shaw DE (2011) How Fast-Folding Proteins Fold. Science 334: 517–520.

[11] Andersen HC (1980) Molecular dynamics simulations at constant pressure and/or temperature. Journal of Chemical Physics 72: 2384–2393.

[12] Berendsen HJC, Postma JPM, Gunsteren WFv, DiNola A, Haak JR (1984) Molecular dynamics with coupling to an external bath. The Journal of Chemical Physics 81: 3684–3690.

[13] Nos S (1984) A unified formulation of the constant temperature molecular dynamics methods. The Journal of Chemical Physics 81: 511–519.

[14] Hoover WG (1985) Canonical dynamics: Equilibrium phase-space distributions. Physical Review A 31: 1695–1697.

[15] Martyna GJ, Tobias DJ, Klein ML (1994) Constant pressure molecular dynamics algorithms. The Journal of Chemical Physics 101: 4177–4189.

[16] Best RB, Hummer G (2009) Optimized Molecular Dynamics Force Fields Applied to the HelixCoil Transition of Polypeptides. The Journal of Physical Chemistry B 113: 9004–9015.

[17] Lindorff-Larsen K, Piana S, Palmo K, Maragakis P, Klepeis JL, et al. (2010) Improved side-chain torsion potentials for the Amber ff99sb protein force field. Proteins 78: 1950–1958.

[18] Li DW, Bruschweiler R (2011) Iterative Optimization of Molecular Mechanics Force Fields from NMR Data of Full-Length Proteins. Journal of Chemical Theory and Computation 7: 1773–1782.

[19] Best RB, Zhu X, Shim J, Lopes PEM, Mittal J, et al. (2012) Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone , and side-chain (1) and (2) dihedral angles. Journal of Chemical Theory and Computation 8: 3257–3273.

[20] Piana S, Lindorff-Larsen K, Dirks RM, Salmon JK, Dror RO, et al. (2012) Evaluating the Effects of Cutoffs and Treatment of Long-range Electrostatics in Protein Folding Simulations. PLoS ONE 7: e39918.

[21] Darden T, York D, Pedersen L (1993) Particle mesh Ewald: An Nlog(N) method for Ewald sums in large systems. The Journal of Chemical Physics 98: 10089–10092.

[22] Siggers K (2004) Loop dynamics of fibronectin type III domain: A study by NMR spectroscopy. Ph.D., Columbia University.

[23] OConnell NE (2009) On the Use of NMR Spin Relaxation Spectroscopy to Characterize Conformational Dynamics in Proteins. Ph.D., Columbia University.

[24] Gardner KH, Kay LE (1998) The use of 2h, 13c, 15n multidimensional NMR to study the structure and dynamics of proteins. Annual Review of Biophysics and Biomolecular Structure 27: 357–406.

[25] Goto NK, Gardner KH, Mueller GA, Willis RC, Kay LE (1999) A robust and cost-effective method for the production of Val, Leu, Ile (delta 1) methyl-protonated 15n-, 13c-, 2h-labeled proteins. Journal of biomolecular NMR 13: 369–374.

[26] Goto NK, Kay LE (2000) New developments in isotope labeling strategies for protein solution NMR spectroscopy. Current Opinion in Structural Biology 10: 585–592.

[27] Tugarinov V, Kay LE (2004) An isotope labeling strategy for methyl TROSY spectroscopy. Journal of biomolecular NMR 28: 165–172.

[28] Tugarinov V, Kay LE (2005) Methyl groups as probes of structure and dynamics in NMR studies of high-molecular-weight proteins. Chembiochem: A European Journal of Chemical Biology 6: 1567–1577.

[29] Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, et al. (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. Journal of biomolecular NMR 6: 277–293.

[30] Sklen V, Bax A (1987) Spin-echo water suppression for the generation of pure-phase two-dimensional NMR spectra. Journal of Magnetic Resonance (1969) 74: 469–479.

[31] Kabsch W (1993) Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. Journal of Applied Crystallography 26: 795–800.

[32] Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P, et al. (2011) Overview of the *CCP* 4 suite and current developments. Acta Crystallographica Section D Biological Crystallography 67: 235–242.

[33] McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, et al. (2007) *Phaser* crystallographic software. Journal of Applied Crystallography 40: 658–674.

[34] Zeiske T, Stafford KA, Friesner RA, Palmer AG (2013) Starting-structure dependence of nanosecond timescale intersubstate transitions and reproducibility of MD-derived order parameters. Proteins: Structure, Function, and Bioinformatics 81: 499–509.

[35] Stafford K (2013) Thermal adaptation of conformational dynamics in ribonuclease H. Ph.D., Columbia University.

[36] Stafford KA, Ferrage F, Cho JH, Palmer AG (2013) Side chain dynamics of carboxyl and carbonyl groups in the catalytic function of Escherichia coli ribonuclease H. Journal of the American Chemical Society 135: 18024–18027.

[37] Stafford KA, Palmer AG (2014) Evidence from molecular dynamics simulations of conformational preorganization in the ribonuclease H active site. F1000Research 3: 67.

[38] Stafford KA, Robustelli P, Palmer AG (2013) Thermal adaptation of conformational dynamics in ribonuclease H. PLoS computational biology 9: e1003218.

[39] Stafford KA, Trbovic N, Butterwick JA, Abel R, Friesner RA, et al. (2015) Conformational preferences underlying reduced activity of a thermophilic ribonuclease H. Journal of Molecular Biology 427: 853–866.

[40] Derrick JP, Wigley DB (1994) The third IgG-binding domain from streptococcal protein G. An analysis by X-ray crystallography of the structure alone and in a complex with Fab. Journal of Molecular Biology 243: 906–918.

[41] Trbovic N, Kim B, Friesner RA, Palmer AG (2008) Structural analysis of protein dynamics by MD simulations and NMR spin-relaxation. Proteins 71: 684–694.

[42] Hall JB, Fushman D (2006) Variability of the 15n chemical shielding tensors in the B3 domain of protein G from 15n relaxation measurements at several fields. Implications for backbone order parameters. Journal of the American Chemical Society 128: 7855–7870.

[43] Yao L, Grishaev A, Cornilescu G, Bax A (2010) Site-specific backbone amide (15)N chemical shift anisotropy tensors in a small protein from liquid crystal and cross-correlated relaxation measurements. Journal of the American Chemical Society 132: 4295–4309.

[44] DeLano W. The PyMOL molecular graphics system. URL `http://www.pymol.org`.

[45] Matthew P Jacobson GAK (2002) Force Field Validation Using Protein Side Chain Prediction. Journal of Physical Chemistry B - J PHYS CHEM B 106.

[46] Maestro, Schrödinger, LLC, New York, NY, 2009. URL `http://www.schrodinger.com/Maestro/`.

[47] Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. The Journal of chemical physics 79: 926–935.

[48] Case DA, Darden TA, Cheatham TE, Simmerling CL, Wang J, et al. AMBER 9; University of California, San Francisco.

[49] Joshi R, Passner JM, Rohs R, Jain R, Sosinsky A, et al. (2007) Functional specificity of a Hox protein mediated by the recognition of minor groove structure. Cell 131: 530–543.

[50] Bowers KJ, Chow E, Xu H, Dror RO, Eastwood MP, et al. (2006) Scalable algorithms for molecular dynamics simulations on commodity clusters. In: Proceedings of the 2006 ACM/IEEE conference on Supercomputing. Tampa, Florida: ACM, p. 84. doi: 10.1145/1188455.1188544.

[51] Krutler V, van Gunsteren WF, Hnenberger PH (2001) A fast SHAKE algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. Journal of Computational Chemistry 22: 501–508.

[52] Humphrey W, Dalke A, Schulten K (1996) VMD: visual molecular dynamics. Journal of Molecular Graphics 14: 33–38, 27–28.

[53] Chandrasekhar I, Clore GM, Szabo A, Gronenborn AM, Brooks BR (1992) A 500 ps molecular dynamics simulation study of interleukin-1 beta in water. Correlation with nuclear magnetic resonance spectroscopy and crystallography. Journal of Molecular Biology 226: 239–250.

[54] Case DA (1999) Calculations of NMR dipolar coupling strengths in model peptides. Journal of biomolecular NMR 15: 95–102.

[55] Lipari G, Szabo A (1982) Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. Theory and range of validity. Journal of the American Chemical Society 104: 4546–4559.

[56] Jarymowycz VA, Stone MJ (2006) Fast time scale dynamics of protein backbones: NMR relaxation methods, applications, and functional consequences. Chemical Reviews 106: 1624–1671.

[57] Meiler J, Prompers JJ, Peti W, Griesinger C, Brschweiler R (2001) Model-free approach to the dynamic interpretation of residual dipolar couplings in globular proteins. Journal of the American Chemical Society 123: 6098–6107.

[58] Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, et al. (2006) Comparison of multiple Amber force fields and development of improved protein backbone parameters. Proteins 65: 712–725.

[59] Khrov P, De Simone A, Otyepka M, Best RB (2012) Force-field dependence of chignolin folding and misfolding: comparison with experiment and redesign. Biophysical Journal 102: 1897–1906.

[60] Genheden S, Diehl C, Akke M, Ryde U (2010) Starting-Condition Dependence of Order Parameters Derived from Molecular Dynamics Simulations. Journal of Chemical Theory and Computation 6: 2176–2190.

[61] Koller AN, Schwalbe H, Gohlke H (2008) Starting structure dependence of NMR order parameters derived from MD simulations: implications for judging force-field quality. Biophysical Journal 95: L04–06.

[62] Buck M, Bouguet-Bonnet S, Pastor RW, MacKerell AD (2006) Importance of the CMAP correction to the CHARMM22 protein force field: dynamics of hen lysozyme. Biophysical Journal 90: L36–38.

[63] Mackerell AD, Feig M, Brooks CL (2004) Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. Journal of Computational Chemistry 25: 1400–1415.

[64] Best RB, Buchete NV, Hummer G (2008) Are current molecular dynamics force fields too helical? Biophysical Journal 95: L07–09.

[65] Forrest LR, Honig B (2005) An assessment of the accuracy of methods for predicting hydrogen positions in protein structures. Proteins 61: 296–309.

[66] Word JM, Lovell SC, LaBean TH, Taylor HC, Zalis ME, et al. (1999) Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. Journal of Molecular Biology 285: 1711–1733.

[67] Markwick PRL, Bouvignies G, Blackledge M (2007) Exploring multiple timescale motions in protein GB3 using accelerated molecular dynamics and NMR spectroscopy. Journal of the American Chemical Society 129: 4724–4730.

[68] Liwo A, Czaplewski C, Odziej S, Scheraga HA (2008) Computational techniques for efficient conformational sampling of proteins. Current Opinion in Structural Biology 18: 134–139.

[69] Okamoto Y (2004) Generalized-ensemble algorithms: enhanced sampling techniques for Monte Carlo and molecular dynamics simulations. Journal of Molecular Graphics & Modelling 22: 425–439.

[70] Mitsutake A, Sugita Y, Okamoto Y (2001) Generalized-ensemble algorithms for molecular simulations of biopolymers. Biopolymers 60: 96–123.

[71] Chou KC, Carlacci L (1991) Simulated annealing approach to the study of protein structures. Protein Engineering 4: 661–667.

[72] Elber R (2005) Long-timescale simulation methods. Current Opinion in Structural Biology 15: 151–156.

[73] Lei H, Duan Y (2007) Improved sampling methods for molecular simulation. Current Opinion in Structural Biology 17: 187–191.

[74] Bornscheuer UT, Huisman GW, Kazlauskas RJ, Lutz S, Moore JC, et al. (2012) Engineering the third wave of biocatalysis. Nature 485: 185–194.

[75] Zamost BL, Nielsen HK, Starnes RL (1991) Thermostable enzymes for industrial applications. Journal of Industrial Microbiology 8: 71–81.

[76] Kristjansson JK (1989) Thermophilic organisms as sources of thermostable enzymes. Trends in Biotechnology 7: 349–353.

[77] Yang JS, Wallin S, Shakhnovich EI (2008) Universality and diversity of folding mechanics for three-helix bundle proteins. Proceedings of the National Academy of Sciences of the United States of America 105: 895–900.

[78] Piana S, Lindorff-Larsen K, Shaw DE (2013) Atomic-level description of ubiquitin folding. Proceedings of the National Academy of Sciences of the United States of America 110: 5915–5920.

[79] Khan S, Vihinen M (2010) Performance of protein stability predictors. Human Mutation 31: 675–684.

[80] Thiltgen G, Goldstein RA (2012) Assessing predictors of changes in protein stability upon mutation using self-consistency. PloS One 7: e46084.

[81] Potapov V, Cohen M, Schreiber G (2009) Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. Protein engineering, design & selection: PEDS 22: 553–560.

[82] Tian J, Woodard JC, Whitney A, Shakhnovich EI (2015) Thermal stabilization of dihydrofolate reductase using monte carlo unfolding simulations and its functional consequences. PLoS computational biology 11: e1004207.

[83] Wijma HJ, Floor RJ, Jekel PA, Baker D, Marrink SJ, et al. (2014) Computationally designed libraries for rapid enzyme stabilization. Protein Engineering Design and Selection 27: 49–58.

[84] Palmer AG (2015) Enzyme dynamics from NMR spectroscopy. Accounts of Chemical Research 48: 457–465.

[85] Butterwick JA, Palmer AG (2006) An inserted Gly residue fine tunes dynamics between mesophilic and thermophilic ribonucleases H. Protein Science: A Publication of the Protein Society 15: 2697–2707.

[86] Butterwick JA, Loria JP, Astrof NS, Kroenke CD, Cole R, et al. (2004) Multiple time scale backbone dynamics of homologous thermophilic and mesophilic ribonuclease HI enzymes. Journal of Molecular Biology 339: 855–871.

[87] Kanaya S (1998) Enzymic activity and protein stability of E. coli ribonuclease HI. In: Crouch R, Toulm J, editors, Ribonucleases H, Paris: INSERM. pp. 1–38.

[88] Kroenke CD, Loria JP, Lee LK, Rance M, Palmer AG (1998) Longitudinal and Transverse 1h15n Dipolar/15n Chemical Shift Anisotropy Relaxation Interference: Unambiguous Determination of Rotational Diffusion Tensors and Chemical Exchange Effects in Biological Macromolecules. Journal of the American Chemical Society 120: 7905–7915.

[89] Kroenke CD, Rance M, Palmer AG (1999) Variability of the 15n Chemical Shift Anisotropy in Escherichia coli Ribonuclease H in Solution. Journal of the American Chemical Society 121: 10119–10125.

[90] Mandel AM, Akke M, Palmer AG (1995) Backbone dynamics of Escherichia coli ribonuclease HI: correlations with structure and function in an active enzyme. Journal of Molecular Biology 246: 144–163.

[91] Mandel AM, Akke M, Palmer AG (1996) Dynamics of ribonuclease H: temperature dependence of motions on multiple time scales. Biochemistry 35: 16009–16023.

[92] Hollien J, Marqusee S (1999) Structural distribution of stability in a thermophilic enzyme. Proceedings of the National Academy of Sciences of the United States of America 96: 13674–13678.

[93] Hollien J, Marqusee S (1999) A Thermodynamic Comparison of Mesophilic and Thermophilic Ribonucleases H. Biochemistry 38: 3831–3836.

[94] Kanaya S, Itaya M (1992) Expression, purification, and characterization of a recombinant ribonuclease H from Thermus thermophilus HB8. The Journal of Biological Chemistry 267: 10184–10192.

[95] Hostomsky Z, Hostomska Z, Mathews D (1993) Ribonucleases H. In: Linn S, Roberts R, editors, Nucleases, Cold Spring Harbor Laboratory Press. 2nd edition, pp. 341–76.

[96] Cerritelli SM, Crouch RJ (2009) Ribonuclease H: the enzymes in eukaryotes. FEBS Journal 276: 1494–1505.

[97] Tadokoro T, You DJ, Abe Y, Chon H, Matsumura H, et al. (2007) Structural, thermodynamic, and mutational analyses of a psychrotrophic RNase HI. Biochemistry 46: 7460–7468.

[98] Ratcliff K, Corn J, Marqusee S (2009) Structure, stability, and folding of ribonuclease H1 from the moderately thermophilic Chlorobium tepidum: comparison with thermophilic and mesophilic homologues. Biochemistry 48: 5890–5898.

[99] Maragakis P, Lindorff-Larsen K, Eastwood MP, Dror RO, Klepeis JL, et al. (2008) Microsecond Molecular Dynamics Simulation Shows Effect of Slow Loop Dynamics on Backbone Amide Order Parameters of Proteins. The Journal of Physical Chemistry B 112: 6155–6158.

[100] Vugmeyster L, Trott O, James McKnight C, Raleigh DP, Palmer AG (2002) Temperature-dependent Dynamics of the Villin Headpiece Helical Subdomain, An Unusually Small Thermostable Protein. Journal of Molecular Biology 320: 841–854.

[101] Massi F, Palmer AG (2003) Temperature dependence of NMR order parameters and protein dynamics. Journal of the American Chemical Society 125: 11158–11159.

[102] Johnson E, Palmer AG, Rance M (2007) Temperature dependence of the NMR generalized order parameter. Proteins 66: 796–803.

[103] Ishikawa K, Nakamura H, Morikawa K, Kimura S, Kanaya S (1993) Cooperative stabilization of Escherichia coli ribonuclease HI by insertion of Gly-80b and Gly-77–>Ala substitution. Biochemistry 32: 7136–7142.

[104] Ishikawa K, Kimura S, Kanaya S, Morikawa K, Nakamura H (1993) Structural study of mutants of Escherichia coli ribonuclease HI with enhanced thermostability. Protein Engineering 6: 85–91.

[105] Nowotny M, Gaidamakov SA, Ghirlando R, Cerritelli SM, Crouch RJ, et al. (2007) Structure of human RNase H1 complexed with an RNA/DNA hybrid: insight into HIV reverse transcription. Molecular Cell 28: 264–276.

[106] Kimura S, Nakamura H, Hashimoto T, Oobatake M, Kanaya S (1992) Stabilization of Escherichia coli ribonuclease HI by strategic replacement of amino acid residues with those from the thermophilic counterpart. The Journal of Biological Chemistry 267: 21535–21542.

[107] Garcia-Fernndez J (2005) Hox, ParaHox, ProtoHox: facts and guesses. Heredity 94: 145–152.

[108] Passner JM, Ryoo HD, Shen L, Mann RS, Aggarwal AK (1999) Structure of a DNA-bound Ultrabithorax-Extradenticle homeodomain complex. Nature 397: 714–719.

[109] Pearson JC, Lemons D, McGinnis W (2005) Modulating Hox gene functions during animal body patterning. Nature Reviews Genetics 6: 893–904.

[110] Panzer S, Weigel D, Beckendorf SK (1992) Organogenesis in Drosophila melanogaster: embryonic salivary gland determination is controlled by homeotic and dorsoventral patterning genes. Development (Cambridge, England) 114: 49–57.

[111] Yao LC, Liaw GJ, Pai CY, Sun YH (1999) A common mechanism for antenna-to-Leg transformation in Drosophila: suppression of homothorax transcription by four HOM-C genes. Developmental Biology 211: 268–276.

[112] Gehring WJ, Qian YQ, Billeter M, Furukubo-Tokunaga K, Schier AF, et al. (1994) Homeodomain-DNA recognition. Cell 78: 211–223.

[113] Mann RS, Chan SK (1996) Extra specificity from extradenticle: the partnership between HOX and PBX/EXD homeodomain proteins. Trends in genetics: TIG 12: 258–262.

[114] Mann RS, Affolter M (1998) Hox proteins meet more partners. Current Opinion in Genetics & Development 8: 423–429.

[115] Ryoo HD, Mann RS (1999) The control of trunk Hox specificity and activity by Extradenticle. Genes & Development 13: 1704–1716.

[116] Mann RS (1995) The specificity of homeotic gene function. BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology 17: 855–863.

[117] Piper DE, Batchelor AH, Chang CP, Cleary ML, Wolberger C (1999) Structure of a HoxB1-Pbx1 heterodimer bound to DNA: role of the hexapeptide and a fourth homeodomain helix in complex formation. Cell 96: 587–597.

[118] Abe N, Dror I, Yang L, Slattery M, Zhou T, et al. (2015) Deconvolving the recognition of DNA shape from sequence. Cell 161: 307–318.

[119] Buchan DWA, Minneci F, Nugent TCO, Bryson K, Jones DT (2013) Scalable web services for the PSIPRED Protein Analysis Workbench. Nucleic Acids Research 41: W349–357.

[120] Slattery M, Riley T, Liu P, Abe N, Gomez-Alcala P, et al. (2011) Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. Cell 147: 1270–1282.

[121] Gebe JA, Delrow JJ, Heath PJ, Fujimoto BS, Stewart DW, et al. (1996) Effects of Na+ and Mg2+ on the structures of supercoiled DNAs: comparison of simulations with experiments. Journal of Molecular Biology 262: 105–128.

[122] Han B, Liu Y, Ginzinger SW, Wishart DS (2011) SHIFTX2: significantly improved protein chemical shift prediction. Journal of Biomolecular Nmr 50: 43–57.

[123] Lu Q, Knoepfler PS, Scheele J, Wright DD, Kamps MP (1995) Both Pbx1 and E2a-Pbx1 bind the DNA motif ATCAATCAA cooperatively with the products of multiple murine Hox genes, some of which are themselves oncogenes. Molecular and Cellular Biology 15: 3786–3795.

[124] Chang CP, Brocchieri L, Shen WF, Largman C, Cleary ML (1996) Pbx modulation of Hox homeodomain amino-terminal arms establishes different DNA-binding specificities across the Hox locus. Molecular and Cellular Biology 16: 1734–1745.

[125] Berger MF, Badis G, Gehrke AR, Talukder S, Philippakis AA, et al. (2008) Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. Cell 133: 1266–1276.

[126] Noyes MB, Christensen RG, Wakabayashi A, Stormo GD, Brodsky MH, et al. (2008) Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. Cell 133: 1277–1289.

[127] Chan SK, Jaffe L, Capovilla M, Botas J, Mann RS (1994) The DNA binding specificity of Ultrabithorax is modulated by cooperative interactions with extradenticle, another homeoprotein. Cell 78: 603–615.

[128] Lu Q, Kamps MP (1997) Heterodimerization of Hox proteins with Pbx1 and oncoprotein E2a-Pbx1 generates unique DNA-binding specifities at nucleotides predicted to contact the N-terminal arm of the Hox homeodomain–demonstration of Hox-dependent targeting of E2a-Pbx1 in vivo. Oncogene 14: 75–83.

[129] LaRonde-LeBlanc NA, Wolberger C (2003) Structure of HoxA9 and Pbx1 bound to DNA: Hox hexapeptide and DNA recognition anterior to posterior. Genes & Development 17: 2060–2072.

[130] Cribbs DL, Benassayag C, Randazzo FM, Kaufman TC (1995) Levels of homeotic protein function can determine developmental identity: evidence from low-level expression of the Drosophila homeotic gene proboscipedia under Hsp70 control. The EMBO journal 14: 767–778.

[131] Duboule D (1991) Patterning in the vertebrate limb. Current Opinion in Genetics and Development 1: 211–216.

[132] Zhang Y, Larsen CA, Stadler HS, Ames JB (2011) Structural basis for sequence specific DNA binding and protein dimerization of HOXA13. PloS One 6: e23069.

[133] Foos N, Maurel-Zaffran C, Mat MJ, Vincentelli R, Hainaut M, et al. (2015) A flexible extension of the Drosophila ultrabithorax homeodomain defines a novel Hox/PBC interaction mode. Structure (London, England: 1993) 23: 270–279.

[134] Luscombe NM, Laskowski RA, Thornton JM (1997) NUCPLOT: A Program to Generate Schematic Diagrams of Protein-Nucleic Acid Interactions. Nucleic Acids Research 25: 4940–4945.

[135] Billeter M, Qian YQ, Otting G, Mller M, Gehring W, et al. (1993) Determination of the nuclear magnetic resonance solution structure of an Antennapedia homeodomain-DNA complex. Journal of Molecular Biology 234: 1084–1093.

[136] Fraenkel E, Rould MA, Chambers KA, Pabo CO (1998) Engrailed homeodomain-DNA complex at 2.2 resolution: a detailed view of the interface and comparison with other engrailed structures1. Journal of Molecular Biology 284: 351–361.

[137] Berezovsky IN, Chen WW, Choi PJ, Shakhnovich EI (2005) Entropic Stabilization of Proteins and Its Proteomic Consequences. PLoS Comput Biol 1: e47.

[138] Rohs R, West SM, Sosinsky A, Liu P, Mann RS, et al. (2009) The role of DNA shape in protein-DNA recognition. Nature 461: 1248–1253.

[139] Zhou T, Yang L, Lu Y, Dror I, Dantas Machado AC, et al. (2013) DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. Nucleic Acids Research 41: W56–62.

[140] Blanchet C, Pasi M, Zakrzewska K, Lavery R (2011) CURVES+ web server for analyzing and visualizing the helical, backbone and groove parameters of nucleic acid structures. Nucleic Acids Research 39: W68–W73.

[141] Lavery R, Moakher M, Maddocks JH, Petkeviciute D, Zakrzewska K (2009) Conformational analysis of nucleic acids revisited: Curves+. Nucleic Acids Research 37: 5917–5929.

[142] West SM, Rohs R, Mann RS, Honig B (2010) Electrostatic interactions between arginines and the minor groove in the nucleosome. Journal of Biomolecular Structure & Dynamics 27: 861–866.

[143] Rohs R, Jin X, West SM, Joshi R, Honig B, et al. (2010) Origins of specificity in protein-DNA recognition. Annual Review of Biochemistry 79: 233–269.

[144] Bishop EP, Rohs R, Parker SCJ, West SM, Liu P, et al. (2011) A map of minor groove shape and electrostatic potential from hydroxyl radical cleavage patterns of DNA. ACS chemical biology 6: 1314–1320.

[145] Sprules T, Green N, Featherstone M, Gehring K (2003) Lock and key binding of the HOX YPWM peptide to the PBX homeodomain. The Journal of Biological Chemistry 278: 1053–1058.

[146] Sprules T, Green N, Featherstone M, Gehring K (2000) Conformational changes in the PBX homeodomain and C-terminal extension upon binding DNA and HOX-derived YPWM peptides. Biochemistry 39: 9943–9950.

[147] Lu Q, Kamps MP (1996) Structural determinants within Pbx1 that mediate cooperative DNA binding with pentapeptide-containing Hox proteins: proposal for a model of a Pbx1-Hox-DNA complex. Molecular and Cellular Biology 16: 1632–1640.

[148] Green NC, Rambaldi I, Teakles J, Featherstone MS (1998) A Conserved C-terminal Domain in PBX Increases DNA Binding by the PBX Homeodomain and Is Not a Primary Site of Contact for the YPWM Motif of HOXA1. Journal of Biological Chemistry 273: 13273–13279.

[149] Lelli KM, Noro B, Mann RS (2011) Variable motif utilization in homeotic selector (Hox)-cofactor complex formation controls specificity. Proceedings of the National Academy of Sciences of the United States of America 108: 21122–21127.

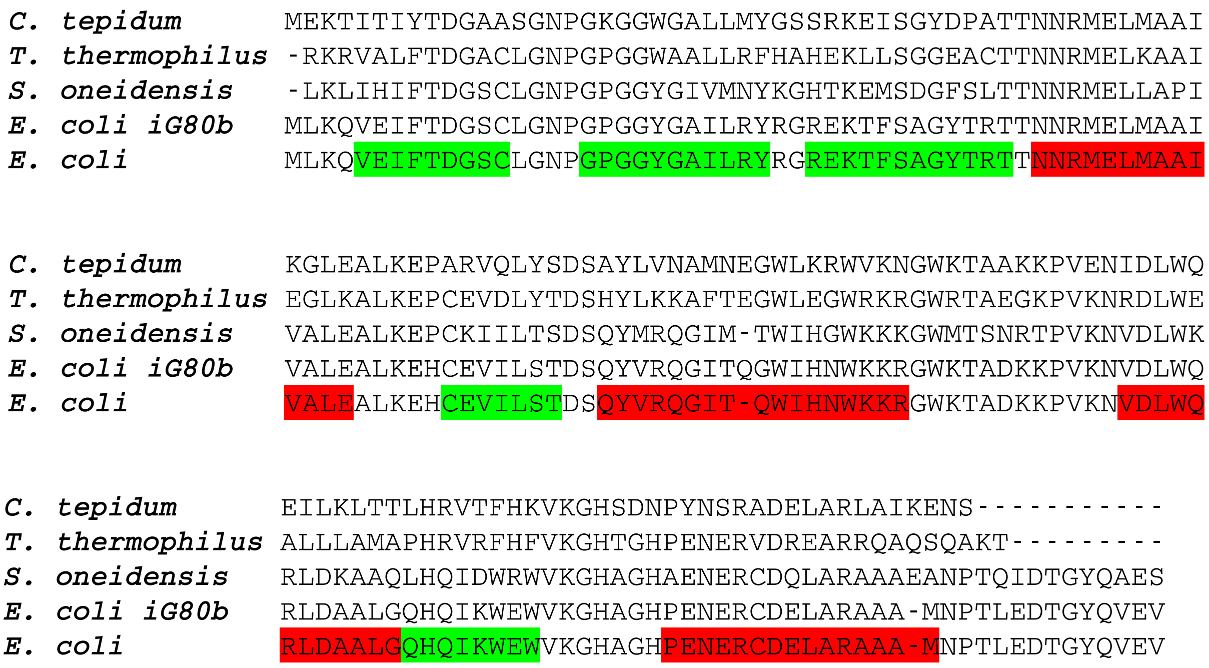
[150] Mann RS, Lelli KM, Joshi R (2009) Hox Specificity: Unique Roles for Cofactors and Collaborators. Current topics in developmental biology 88: 63–101.

[151] Coiffier D, Charroux B, Kerridge S (2008) Common functions of central and posterior Hox genes for the repression of head in the trunk of Drosophila. Development (Cambridge, England) 135: 291–300.

[152] Crocker J, Abe N, Rinaldi L, McGregor AP, Frankel N, et al. (2015) Low affinity binding site clusters confer hox specificity and regulatory robustness. Cell 160: 191–203.

[153] Bostock-Smith CE, Harris SA, Laughton CA, Searle MA (2001) Induced fit DNA recognition by a minor groove binding analogue of Hoechst 33258: fluctuations in DNA A tract structure investigated by NMR and molecular dynamics simulations. Nucleic Acids Research 29: 693–702.

Appendices



```
C. tepidum        MEKTITIYTDGAASGNPGKGGWGALLMYGSSRKEISGYDPATTNNRMELMAAI
T. thermophilus   -RKRVALFTDGACLGNPGPGGWAALLRFHAHEKLLSGGEACTTNNRMELKAAI
S. oneidensis     -LKLIHIFTDGSCLGNPGPGGYGIVMNYKGHTKEMSDGFSLTTNNRMELLAPI
E. coli iG80b     MLKQVEIFTDGSCLGNPGPGGYGAILRYRGREKTFSAGYTRTTNNRMELMAAI
E. coli           MLKQVEIFTDGSCLGNPGPGGYGAILRYRGREKTFSAGYTRTTNNRMELMAAI

C. tepidum        KGLEALKEPARVQLYSDSAYLVNAMNEGWLKRWVKNGWKTAAKKPVENIDLWQ
T. thermophilus   EGLKALKEPCEVDLYTDSHYLKKAFTEGWLEGWRKRGWRTAEGKPVKNRDLWE
S. oneidensis     VALEALKEPCKIILTSDSQYMRQGIM-TWIHGWKKKGWMTSNRTPVKNVDLWK
E. coli iG80b     VALEALKEHCEVILSTDSQYVRQGITQGWIHNWKKRGWKTADKKPVKNVDLWQ
E. coli           VALEALKEHCEVILSTDSQYVRQGIT-QWIHNWKKRGWKTADKKPVKNVDLWQ

C. tepidum        EILKLTTLHRVTFHKVKGHSDNPYNSRADELARLAIKENS-----------
T. thermophilus   ALLLAMAPHRVRFHFVKGHTGHPENERVDREARRQAQSQAKT---------
S. oneidensis     RLDKAAQLHQIDWRWVKGHAGHAENERCDQLARAAAEANPTQIDTGYQAES
E. coli iG80b     RLDAALGQHQIKWEWVKGHAGHPENERCDELARAAA-MNPTLEDTGYQVEV
E. coli           RLDAALGQHQIKWEWVKGHAGHPENERCDELARAAA-MNPTLEDTGYQVEV
```

Figure 1: **Multiple sequence alignment of the five RNases H used herein**
Multiple sequence alignment of all five proteins, with secondary structural elements highlighted for the E. coli sequence (helices in red, strands in light green).