Protein Dynamics

Importance to conformational dissimilarity and use in ligand binding site identification

by

Tommy Carstensen, M.Sc.

The thesis is submitted to University College Dublin for the degree of Ph.D. in the College of Life Sciences

January, 2012

School of Biomolecular and Biomedical Sciences

Head of School: Dr. Gethin McBean

Head of Discipline, Biochemistry: Dr. Margaret Worrall

Supervisor: Dr. Jens Erik Nielsen

1 Table of Contents

2 Acknowledgements

First and foremost I would like to thank my supervisor Jens Erik Nielsen for his advice and guidance in writing this thesis. He has time and again proven himself to be a friend and I am indebted to him. I would like to thank Lynn Ten Eyck for insightful discussions on math related problems, all the members of the Nielsen group for a lot of good memories and discussions at the Conway Institute. In particular I would like to thank Damien Farrell for his computer assistance, Una Bjarnadottir for her infectious laughter, Predrag Kukic for the many gigs we have attended together and Barbara Tynan-Conolly, Fergal O'Meara, Helen Webb, Chresten Rauff Soendergaard and John Bradley for conversations about life and all the little but important things in it. I would like to thank Aniello Palma, Thomas Aidan Digby, Palle Martin Jensen and all my friends in Denmark for keeping in touch despite the distance. I am grateful to Carmina Abengoza, Lourdes Abengoza and Margrethe Hartmann Andersen for their indirect support. I would like to thank Grace Hsin-ju Lin for having been patient and having been a ray of light in a dark period of my life. My sister Elisa Carmen Carstensen I am grateful to for her support, when I needed it the most. Finally I would like to thank my mom Antonia Salamanca Camacho for her continuous support and unconditional love despite her own challenges.

3 Preface

This thesis was produced in the School of Biomolecular and Biomedical Science at University College Dublin under the supervision of Dr. Jens Erik Nielsen. The thesis is centered on the investigation of structural variability of proteins due to their flexibility, due to mutations and due to the method of solving their structure. The thesis examines structural differences and similarities between sequence identical and sequence similar structures. The thesis investigates the relationship between the intrinsic dynamics of proteins and the change in protein conformation upon ligand binding and how this relationship can be used to predict ligand binding sites.

Chapter 1 is the introduction.

Chapter 2 provides a thorough analysis of the structural differences and similarities between a subset of structures in the protein data bank (PDB). Those structures are hen egg white lysozyme (HEWL) and bacteriophage T4 lysozyme (T4L). Chapter 2 investigates the conformational plasticity of T4L and HEWL as displayed in X-ray structures. In chapter 2 it is examined, whether observed structural differences are due to methodology and procedure rather than being actual structural differences. In chapter 2 it is examined, whether mutations have a significant structural effect on T4L beyond structural differences observed between wild type structures.

Chapter 3 expands on the analysis carried out in chapter 2. All sequence similar structures in the protein data bank (PDB) are compared against each other, and the effect of a range of properties - as a source of structural variation between structures – are investigated. A subset of structures containing only single point mutants are investigated with the aim of determining, if the structural effect of a single point mutation propagates through the structure or only has limited localized effects.

Chapter 4 investigates whether conformational changes observed upon ligand binding via X-ray structures are motions caused by the binding of the ligand or an intrinsic property of the protein and a trajectory and conformational landscape explored by the protein independent of the ligand. Concluding that conformational changes are spontaneous, it is further probed in chapter 4, whether the ligand binding site can be identified by perturbing the system and analyzing, which perturbations cause a disruption of the motion between the open and closed conformation of the protein. For this last analysis I use a novel normal mode analysis tool. In chapter 4 I present an NMA method for prediction of ligand binding sites.

4 Summary

The dynamics of proteins have in the past three decades been speculated to be of importance for protein function. Rational engineering of proteins therefore involves having knowledge about their dynamics. In this thesis I use knowledge about the intrinsic dynamics of proteins to identify their ligand binding site using a normal mode analysis based algorithm. I show this algorithm to be successful in its prediction for proteins binding their ligand by conformational selection. I furthermore analyze a large set of structures to identify, which parameters contribute to structural differences and which structural differences are due to intrinsic dynamics and experimental errors. Specifically I find that space group differences is the single most important parameter for explaining differences between X-ray structures of proteins. This highlights the importance of taking into account crystal contacts when comparing computationally predicted structures to experimental target structures. I show that the author of a structure and the starting model used for solving the phase problem of X-ray crystallography by molecular replacement are more important for the final reported structure, than mutations, the presence of ligands and pH differences. I show that structural heterogeneity causes larger structural differences than mutations. In fact mutational effects are very hard to see both in the vicinity and distant from the site of mutation. This highlights the importance of using conformational ensembles for structure based calculations such as pK_a calculations and stability calculations. When performing a calculation on an ensemble of structures, this will reveal the range of values, which are sampled by the protein at its conformational dynamic equilibrium.

5 Publications and presentations

5.1 Peer reviewed publications

2011 – On the development of protein pKa calculation algorithms

Carstensen T, Farrell D, Huang Y, Baker NA, Nielsen JE

Proteins, 79, 3287

2011 – Progress in the prediction of pKa values in proteins

Alexov E, Mehler EL, Baker N, Baptista AM, Huang Y, Milletti F, Nielsen JE, Farrell D, **Carstensen T**, Olsson MH, Shen JK, Warwicker J, Williams S, Word JM

Proteins, 79, 3260

2010 – Capturing, sharing and analysing biophysical data from protein engineering and protein characterization studies

Farrell D, O'Meara F, Johnston M, Bradley J, Søndergaard CR, Georgi N, Webb H, Tynan-Connolly BM, Bjarnadottir U, **Carstensen T**, Nielsen JE

Nucleic Acids Res, 38, e186

2009 – Structural artifacts in protein-ligand X-ray structures: implications for the development of docking scoring functions

Søndergaard CR, Garrett AE, **Carstensen T**, Pollastri G, Nielsen JE

J Med Chem, 52, 5673

2007 – Fractional ^{13}C enrichment of isolated carbons using [1-^{13}C]- or [2-^{13}C]-glucose facilitates the accurate measurement of dynamics at backbone C$^{\alpha}$ and side-chain methyl positions in proteins

Lundström P, Teilum K, **Carstensen T**, Bezsonova I, Wiesner S, Hansen DF, Religa TL, Akke M, Kay LE

J Biomol NMR, 38, 199

5.2 Poster presentations

2009 - Analysing the determinants of protein dynamics using normal mode analysis

Conway Science Festival, UCD, Dublin, Ireland

2008 – PEAT_SA: A program for high-throughput screening of the effect of single point mutations and its application to drug resistant mutations in HIV Protease

Michael Johnston, Chresten Søndergaard, Tommy Carstensen, Jens Nielsen

8th Annual UCD Conway Festival of Research, UCD, Dublin, Ireland

2007 - Normal mode analysis based on an anisotropic network model: prediction of residues important for protein dynamics in T4 lysozyme and cAMP dependent kinase

Conway Science Festival, UCD, Dublin, Ireland

2007 - Detailed Analysis of the Dynamics of cAMP Dependent Protein Kinase using Molecular Dynamics Simulations and Normal Mode Analysis

The 21st Symposium of The Protein Society, Boston, Massachussets

2007 - Prediction of residues important for protein dynamics. A normal mode analysis approach.

Frontiers of NMR in Molecular Biology, Keystone Symposia, Snowbird, Utah

5.3 Oral presentations

2008 – Analysing the determinants of protein dynamics using normal mode analysis

Conway Bioinformatics Seminar Series, UCD, Ireland

2007 – A study of the directionality, kinetics and thermodynamics of conformational changes using computational normal mode analysis and experimental ^{13}C methyl CPMG relaxation dispersion

Invited talk, Lund University, Department of Biophysical Chemistry, Sweden

2007 – Determination of dynamically important residues in proteins by normal mode analysis - An anisotropic Gaussian network model

Invited talk, August Krogh Institute, University of Copenhagen, Denmark

2007 – Determination of dynamically important residues in proteins by normal mode analysis

Invited talk, Danisco / Genencor, Copenhagen, Denmark

6 Introduction

6.1 Structure of the thesis

This thesis contains three result chapters, numbered 7, 8 and 9. In chapter 7 an investigation of the importance of author, lab and starting model to protein structure is carried out. When comparing protein structures this is something, which is never considered. Here I show those factors to be more important to the final representation of the electron density in coordinate space than physiochemical parameters and even mutations. Chapter 8 deals with the analysis of the structural effect of physiochemical and non-physiochemical properties on protein structures. It is shown that experimental errors and protein dynamics mostly contribute to structural differences. Conclusions from the first and second results chapters are that experimental errors/differences and/or intrinsic dynamics is mostly responsible for observed structural differences.

In the third results chapter it will be described how normal mode analysis has been used to study the change in directions, amplitudes and energies of the dynamics of a protein upon perturbation of the structure of that protein by mimicking the binding of a ligand. For a set of proteins it is investigated whether conformational sampling or ligand induced fit is responsible for the conformational changes observed upon ligand binding. In those cases where conformational selection is predicted to be responsible for conformational changes, the intrinsic dynamics is calculated with normal mode analysis and used to predict ligand binding sites. In summary I present a method for identifying ligand binding sites in proteins governed by spontaneous conformational changes. I also show that the success rate of the ligand binding site finding algorithm is low, when the ligand induces a non-spontaneous conformational change. Most ligand binding site finding algorithms are grid based methods, which do not take conformational change into consideration. Identification of ligand binding sites has been pursued for many decades. Many proteins in the sequence space also exist in the structure space. Therefore the most recent methods of ligand binding site identification involve sequential comparison to already existing structures with known binding sites. My method however is solely structure based.

6.2 Proteins

Proteins are dead end molecules. They will eventually break down and unlike DNA they cannot replicate themselves. The synthesis of proteins is dependent on the “blueprint molecule” DNA(Crick 1958) and humans cannot even synthesize all of the naturally occurring amino acids; i.e. the essential amino acids. Nevertheless proteins are very interesting biological macromolecules, as they are the molecular machinery of the cell. Proteins carry out catalysis and serve structural and signaling purposes. DNA is restricted to a 4 letter alphabet and is at the molecular level restricted to attaining a helix structure. Proteins on the other hand have more than 20 different building blocks at their disposal and despite being in a linear sequence like the nucleotides of DNA they interact to form thousands of different folds. It is the structure of proteins that give them their unique function.

"Almost all aspects of life are engineered at the molecular level, and without understanding molecules we can only have a sketchy understanding of life itself." - Francis Crick 1988 (Crick 1988), p. 61)

6.3 Protein Structure

Before digging deeper into protein dynamics, it is beneficial to have an understanding of protein structure. I dedicate this section to a short walkthrough of protein structures.

A typical protein is a linear sequence of tens to hundreds and even a thousand amino acids. The typical protein only contains the 20 common amino acids found in most bacteria, plants and animals. The 20 amino acids have different properties, because they are made up of different atoms. Because the 20 amino acids have different properties in terms of size, charge and many other parameters, they can be combined in a linear sequence to yield a unique three dimensional structure, which in some cases can automatically fold in solution given the unique sequence of amino acids.(Lumry and Eyring 1954; Haber and Anfinsen 1962; Epstein, Goldberger et al. 1963) The structure that the protein attains is dependent on the interaction between the residues and the surrounding environment. Properties of the solvent – in which the proteins reside – such as polarity, pH, salinity can be changed and in turn change the solubility of the proteins in solution and the conformational distribution of the protein structures.

6.3.1 Protein structure properties

Throughout this thesis I will be referring to various protein structure properties. I feel it is beneficial to the reader to have these metrics and attributes presented here.

One structural property of proteins that I will refer to in chapters 7 and 8 is the φ and ψ dihedral angles, also known as the Ramachandran angles. Those are the dihedral angles for each residue between backbone atoms C_{i-1}, N_i, CA_i, C_i and N_i, CA_i. C_i, N_{i+1}, respectively. It is important to understand that the Ramachandran angles are not randomly distributed, but rather cluster in specific regions of the Ramachandran plot (a plot of ϕ values on one axis and ψ values on the other axis). The cluster positions in the plot are dependent on the type of residue and the type of secondary structure element in which each residue is located (Figure 1). Plotting the Ramachandran angles reveals that the ψ angle is often zero. This is evidence that a final X-ray structure is very dependent on automated methods, just like it is evidence that automated methods were used, when B-factors are all identical or structures solved by molecular replacement change very little from their parent structure. In chapter 7 I present a thorough analysis of structural differences between sequence similar proteins and show that structure similarity is very dependent on molecular replacement among other things.

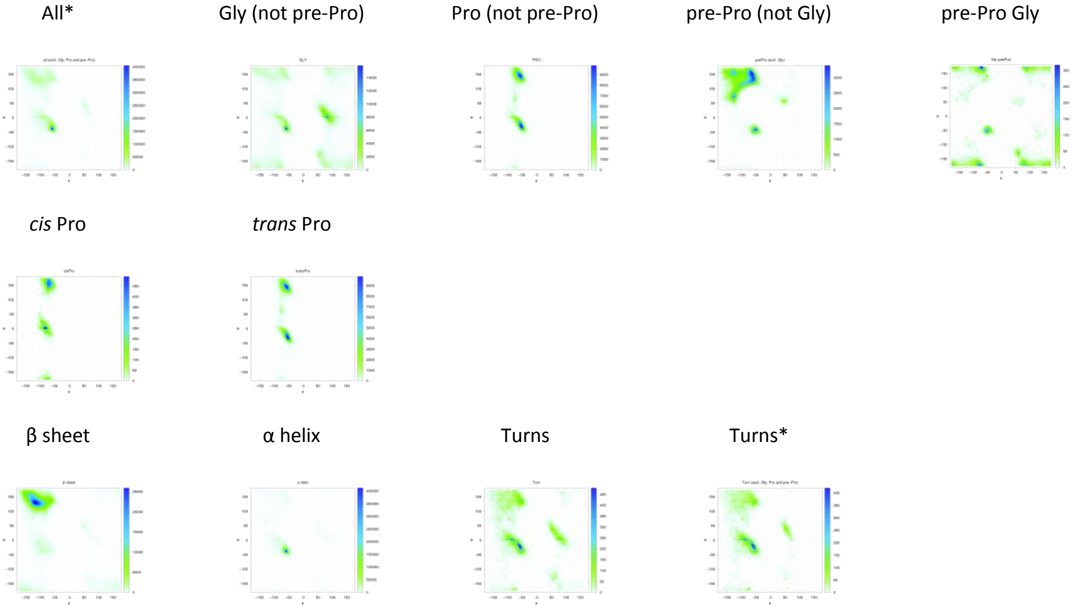


Figure 1 – The Ramachandran plot of different subsets of residue types (top rows) and secondary structure elements (bottom row). All* and Turns* is all residues and residues in turns excl. Gly, Pro, pre-Pro. Glycine frequently resides at positive ϕ angles unlike the other 19 amino acids. The freedom of the dihedral angles of Pro is very limited. Residues just prior to Proline (pre-Pro) have unique angles. Those of Glycine preceding Proline are different from the 19 other amino acids preceding Proline and different from Glycine not preceding Proline. The angles of pre-Pro Gly are different from Glycine residues not followed by Proline residues. Depending on whether Proline is in a *cis* or *trans* configuration it will have slightly different Ramachandran dihedrals. In terms of secondary structure the residues in α helices and β sheets are restricted to a very small area on the Ramachandran plot. A larger variety is seen for residues in turns, although much of the variety can be attributed to Glycine and Proline cf. the comparison between "Turns" and "Turns*". As can be seen from all the φ,ψ plots the ψ angle is frequently 0. This is most likely not a physical phenomenon. Rather it is a sign that many ψ angles are automatically set by crystallographic software packages rather than manually by crystallographers. Figure was drawn with gnuplot 4.4.0.

6.4 Protein Dynamics

Proteins are not rigid structures and their function is dependent on more than the scaffold which their structure provides. Just as a static photo of a clock or a steam engine does not reveal the full picture of their functions, the structure of a protein standing alone does not reveal the details of the function of this biological machine. The functions of proteins can only be fully explained if their dynamic behavior is taken into consideration. The idea of dynamics as an important component of protein function is not a new one. It dates back as far back as the seventies when McCammon and Karplus stated that "a complete description of an enzyme requires a knowledge of its structure and the dynamics of its function."(McCammon, Gelin et al. 1976) Even in 1963, Feynman during his famous Caltech lectures noted that "everything that living things do can be understood in terms of the jiggling and wiggling of atoms". Moreover the CASP members, which do protein structure prediction, acknowledge that intrinsic dynamics is something not easily captured with a single structure and argues that ensembles of structures rather than single structures should be compared with averaged experimental data.(MacCallum, Hua et al. 2009) While the dynamic properties of a protein stem from its structure, today it is widely accepted that dynamics, in addition to structure, is of prime importance to protein function.

There are many indications that dynamics are of importance to enzyme activity and specificity. Here I list a range of examples.

It has been shown that a mutation of two residues in dihydrofolate reductase (DHFR) located far from each other and far from the active site of the enzyme changes the enzyme activity.(Rajagopalan, Lutz et al. 2002; Wang, Goodey et al. 2006) Because the mutations are located far from the active site, the substrate and cofactor binding was not affected, but the mutated residues were shown through experimental NMR methods and computational MD simulations to be part of a dynamic network. Furthermore the double mutation (G121V, M42W) in the same protein - despite the distance from the active site - has been shown to have an effect on the hydride transfer step as measured by the kinetic isotope effect (KIE) after isotope labeling of the NADPH cofactor, which is involved in the hydrogen bond chemistry of the catalytic reaction of DHFR. The KIE was interpreted as a coupling of vibrational energy - involving the sites of mutation far from the active site - coupled to the activation of the catalyzed bond.(Wang, Goodey et al. 2006) More recently it has been shown that the hydride transfer of the DHFR mutant N23PP/S148A is slower than that of the mutant and – while the drop in the catalytic rate and the hydride transfer rate is not proportional to the drop in the rate of the millisecond time scale dynamics in the active site loops – the double mutation does impair the dynamics as measured by CPMG relaxation.(Bhabha, Lee et al. 2011) This is interesting, because the hydride transfer is the chemical step of the catalysis, which directly links conformational fluctuations and catalysis. Electrostatic pre-organization it has been argued is the explanation of the catalytic efficiency of enzymes(Pisliakov, Cao et al. 2009), but for the DHFR double mutant it was however shown, that the pK_a value of the catalytic Met20 residue was unchanged relative to the wt, and the structure of the double mutant was also nearly identical to that of the *wt* (which most mutants are, as I will show in this thesis).

An enzyme (*cis*–*trans* Proline isomerase, cyclophilin A) has been shown to have correlated motions on a time scale identical to the time scale of the catalytic turnover rate in the absence of substrate(Eisenmesser, Millet et al. 2005), which suggests dynamics to be an intrinsic property of the structure. For the same enzyme it has been shown, that a single point mutation causes a proportional drop in the rate of catalysis and the rate of conformational change as measured by NMR, which further suggest a strong link between enzyme catalysis and intrinsic dynamics.(Fraser, Clarkson et al. 2009)

It has been suggested that conformational changes can slow down the binding of ligands (conformational gating).(McCammon and Northrup 1981) It has been further shown that the substrate selectivity of acetylcholine esterase (AChE) is controlled by conformational gating, which changes the probability of the ligand binding site being accessible.(Zhou, Wlodek et al. 1998) The ligand binding site of AChE is buried and the otherwise rapid conformational change between an open and closed state slows down the substrate binding. The substrate binding is slowed the most for large ligands, and protein dynamics (the frequency of the conformational gating) thereby control enzyme specificity. This could maybe explain the reduced $k_{\mathrm{cat}}/K_{\mathrm{m}}$ observed for the hydrolysis of the bulkier butyrylcholine.(Zhou, Wlodek et al. 1998)

Dynamics, as defined in this thesis, is not the molecular motion associated with a chemical reaction (e.g. ATP hydrolysis in the F1-ATPase, muscle filament, ribosome, etc.); rather it is the spontaneous movements driven by thermal energy that are restricted by the backbone framework of the protein and the electrostatic forces between atoms. I refer to these movements as equilibrium fluctuations. In particular I am interested in concerted inter-domain motions, as these are the ones described by low energy normal modes and speculated to be of importance to enzyme catalysis.(Eisenmesser, Bosco et al. 2002; Wolf-Watz, Thai et al. 2004; Eisenmesser, Millet et al. 2005; Henzler-Wildman, Lei et al. 2007; Henzler-Wildman, Thai et al. 2007; Fraser, Clarkson et al. 2009) These inter domain motions however are not suspected to be important for protein folding. The focus of this thesis is on enzyme catalysis rather than protein folding. One could imagine a protein exerting a mechanical force on its substrate during catalysis. This mechanical force could strain the protein into a specific conformation and reduce the free activation energy of the enzyme.(Bustamante, Chemla et al. 2004) The turnover rate would thus be dependent on the intrinsic dynamics of the protein.(Henzler-Wildman, Lei et al. 2007)

I have provided no examples of protein dynamics actively assisting in enzyme catalysis. So far there has been no evidence of this(Pisliakov, Cao et al. 2009), and I do not think protein dynamics is the key to explain catalytic rate enhancement the way it in some cases can explain substrate specificity and rate limitation.

Just as structure and electrostatics can be engineered to change properties of an enzyme such as activity, substrate selectivity, reaction specificity, stability, etc., it is also likely that dynamics can be engineered to optimize enzyme activity. For example, optimization could involve faster substrate binding and/or product release, if any of the steps were shown to be rate limiting to catalysis. The optimization might be carried out by changing the thermodynamics – i.e. the population states – of the reaction towards a transition state like conformation, which favors catalysis. If the population of conformation similar or nearly identical to the transition state is sparsely populated, then the catalytic rate should be able to be increased by further populating the transition state like conformation. Optimization might also be achieved by simply changing the speed at which the functionally important motion occurs. It is this latter approach on which this thesis will focus in particular. The incentive to study dynamics therefore is to be able to predict changes in dynamics upon mutations and the associated changes in activity if any. But mutational changes will most likely affect more than just the dynamics of a protein. Therefore the results and methods presented in this thesis would have to be considered in combination with other methods to make a qualified prediction of the consequences of point mutation(s). And afterwards the time consuming mutational study and subsequent measurement of relevant physical parameters would have to be carried out in order to validate the methods and improve them for future predictions on other systems.

6.4.1 Theoretical investigation of protein dynamics

Protein dynamics can be studied by different computational methods. Simulation of molecular dynamics(Karplus and McCammon 2002) is the most popular method as judged by the number of publications on the subject, and another method is normal mode analysis. Normal mode analysis can be used to describe experimentally observed functional motions(Wang, Borchardt et al. 2005) and equilibrium fluctuations as observed by for example experimental X-ray temperature factors(Bahar, Atilgan et al. 1997; Haliloglu and Bahar 1999).

In this thesis I make use of NMA. Here I briefly present advantages and disadvantages of NMA and MD. Dynamics of proteins can be addressed computationally by full force field molecular dynamics (MD). The first such application to a biological macromolecule was done in the 1970s.(McCammon, Gelin et al. 1977) The disadvantage of MD simulations is the iterative nature of the method, which reduces accuracy after many time steps if a sufficiently small time step is not used. A time step on the fs time scale is often applied and simulations running longer than 100ps are thus time consuming. And only local vibrations can be observed on the ps time scale, whereas secondary structure motions can be on the µs time scale, domain motions and enzymatic catalytic rate constants (k_{cat} values) can be on the ms time scale and protein folding can be on the scale of seconds. Another disadvantage of MD simulations is the dependency on the protein under study and the force field used.

Normal mode analysis on the other hand is based on approximated harmonic equations, which is why the method according to the theory is not suitable for describing transitions between conformations separated by energy barriers. NMA only describes the most likely trajectory of the protein from a given equilibrium position. It does not describe the path between two conformations, if they are separated by an energy barrier. In this thesis it will be shown that NMA actually succeeds very well at describing conformational transitions, which implies a low energy barrier or a directional overlap between initial thermal/spontaneous motion and energy/ligand driven motion. Conformational transitions are even well described when simplifying both the force field in use and the structural/coordinate input. If a uniform force constant is used, it should be noted that normal mode analysis does not reveal information about absolute amplitudes and timescales of motion. Another inaccuracy of normal mode analysis is the fact that solvent is not taken into consideration even though the solvent is expected to be of importance and largely determining the magnitude of protein fluctuations at 180K and above.(Vitkup, Ringe et al. 2000) To leave out the effect of solvents is crucial, if protein processes are truly solvent slaved as proposed.(Fenimore, Frauenfelder et al. 2002) Another drawback is that normal modes describe movements that are linear which is rarely the case for biological motions. In addition linear trajectories are often not sterically allowed.(Flores, Echols et al. 2006) A new NMA method has been developed, which calculates changes to dihedral angles instead of coordinates in Cartesian space.(Bray, Weiss et al. 2011)

The major reason I chose normal mode analysis and not MD simulations to study the dynamics of proteins is that the method is extremely fast compared to MD simulations. Computations and the input for them can be simplified to make the method even faster as will be shown. Because of the procedure being fast it is suitable for studying the effects of perturbations involving individual residues. Another reason for the choice of NMA is that trajectories from normal mode analysis unlike those from MD simulations do not contain noise in the form of side chain motions, which are not of importance for e.g. larger protein domain movements. It is also ambiguous for how long an MD simulation has to run before it yields the functional dynamics of the protein of interest. On the contrary the calculation of normal modes from a given equilibrium position is a finite process.

Similar results can be achieved with NMA and more computationally time consuming MD simulations. Similar to what can be achieved with NMA, MD simulations have been shown to correlate with B-factors of T4L and with fluctuations between ensembles of X-ray structures of T4L.(de Groot, Hayward et al. 1998) It has been shown for a large set of proteins that their essential dynamics calculated from molecular dynamics (i.e. directions and amplitudes of calculated motion) correlate well with normal mode analysis.(Ahmed, Villinger et al. 2010)

Nobody has proven a coupling between protein structure and function on one side and dynamics on the other. Experiments of Dorothee Kern have however suggested the importance of protein dynamics and conformational changes preceding enzyme substrate binding to enzyme catalysis.(Eisenmesser, Bosco et al. 2002; Wolf-Watz, Thai et al. 2004; Eisenmesser, Millet et al. 2005; Henzler-Wildman, Thai et al. 2007; Fraser, Clarkson et al. 2009)

Protein dynamics can be studied by several different methods and on several different time scales (Figure 2). Computational methods include normal mode analysis(Go and Go 1976), molecular dynamics simulations(McCammon, Gelin et al. 1977), ensemble generation and principal component analysis on an ensemble of X-ray structures.(Best, Lindorff-Larsen et al. 2006)

6.4.2 Experimental investigation of protein dynamics

A number of experimental methods are available for the study of protein dynamics. As I do not present any wet lab results in this thesis, I will only briefly describe the available methods. Other methods available, that I do not describe, are shown in Figure 2, along with the time scales that they can be used to study. By comparison to Figure 3 it can be seen that, whereas MD simulations are not a viable method for studying protein conformation changes during enzyme catalysis, many NMR experiments cover the time scale of enzyme catalysis from substrate binding to product release.(Palmer III, Kroenke et al. 2001) A method not shown in Figure 2 is FRET, which can be used to examine the frequency of a conformational change that brings two residues into the vicinity of each other.

In terms of structure, recently it has been shown that the "hidden" (i.e. low population) conformational states of proteins can be determined through a combination of computational methods (CS-Rosetta) and NMR experiments that does not involve the determination of internuclear distances.(Hansen, Vallurupalli et al. 2008; Bouvignies, Vallurupalli et al. 2011) This takes structural bioinformatics and structure based calculations to the next level, because multiple conformers have to be considered / are available when doing structure prediction and structure based calculations.(Ashkenazy, Unger et al. 2011)

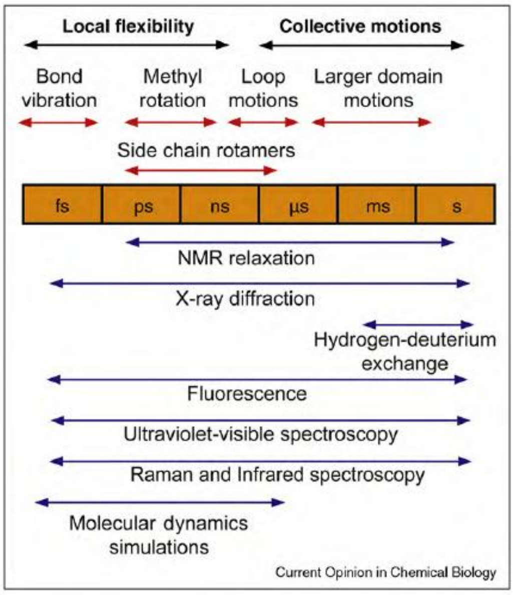


Figure 2 – Time scales of events of protein dynamics (top) and experiments allowing the study of those time scales (bottom). Enzyme catalysis happens on the sub-fs time scale, whereas the large scale concerted conformational changes preceding catalysis happen on the µs-ms time scale. When I refer to protein dynamics in this thesis, it is this latter type that I refer to.(Henzler-Wildman and Kern 2007)

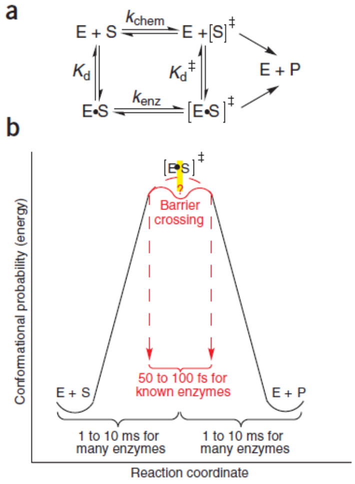

a
E + S
k_{chem}
E + [S]‡
K_d
K_d‡
E + P
E•S
k_{enz}
[E•S]‡
b
[E•S]‡
?
Barrier crossing
50 to 100 fs for known enzymes
E + S
E + P
1 to 10 ms for many enzymes
1 to 10 ms for many enzymes
Conformational probability (energy)
Reaction coordinate

Figure 3 – Time scale of enzymatic reaction and enzyme pre-organization. Dynamics is of importance to the pre-organization of the ligand binding site and the formation of the enzyme-substrate complex rather than the actual enzymatic catalytic step.(Schwartz and Schramm 2009)

6.4.2.1 NMR

NMR is a widespread method to study the dynamics of proteins. Both folding processes(Teilum, Poulsen et al. 2006) and catalytic processes(Eisenmesser, Millet et al. 2005) can be studied with NMR. Experimental methods for qualitatively and quantitatively measuring large- and small-scale dynamics exist. Most interesting in relation to normal mode analysis are methods for quantitative measurement of amplitudes, directions and time scales of movements. Different time scales can be obtained through different NMR relaxation experiments.

As an example CPMG and $R_{1\rho}$ relaxation dispersion experiments are useful for measuring movements approximately on the µs-ms timescale (10^3-$10^6 s^{-1}$). The transverse relaxation is a sum of dipolar relaxation, chemical shift anisotropy relaxation and relaxation due to chemical exchange.(Keeler 2006).

CPMG relaxation dispersion experiments can reveal information about the kinetic (forward and reverse rate constants) and thermodynamic (populations) properties of a conformational change.(Mulder, Hon et al. 2002) CPMG and $R_{1\rho}$ experiments are most suitable for exchange on the $100s^{-1}$-$3000s^{-1}$ and $1000s^{-1}$-$50000s^{-1}$ timescale respectively.(Kempf and Loria 2004), pp. 192) Both are timescales relevant to e.g. protein folding (Fersht 1998), p. 551) and certain protein-protein interactions (Fersht 1998), p. 153).

S^2 order parameters probe the dynamics on the ps-ns timescale (10^{12}-$10^9 s^{-1}$). S^2 order parameters are calculated from longitudinal relaxation rates, R_1, transverse relaxation rates, R_2, and NOEs.(Keeler 2006)

6.4.2.2 B-factors

B-factors are a good gauge of the freedom of movement of individual atoms in a protein structure. I will however cover B-factors more extensively during my walkthrough of the important parameters of X-ray crystallography and the caveats of the method in section 6.5.1.1.3.

6.5 Protein structure determination

To be able to discuss differences between X-ray structures requires a certain level of knowledge about how those structures were achieved experimentally. Here I present some key concepts of X-ray crystallography mentioned throughout the thesis.

6.5.1 Experimental protein structure determination

Without a protein structure, there can be no calculation of the motions of the protein. Protein structures are predominantly determined by X-ray crystallography and NMR spectroscopy. Many macromolecular structures are deposited in the Protein Data Bank(Bernstein, Koetzle et al. 1977; Berman, Westbrook et al. 2000; Henrick, Feng et al. 2008) and made publicly available. Other methods of determining protein structures include homology modeling. If one is merely interested in a coarse grained backbone conformation of a protein and a template structure with a high sequence similarity is available, then homology modeling is an adequate substitute for having an original structure. Since I only work with high resolution X-ray structures and I use details of the diffraction experiment in my analysis of structural differences, it makes sense to give an introduction to X-ray crystallography while skipping methods of NMR used for structure determination.

6.5.1.1 X-ray crystallography

X-ray crystallographic solution of a protein is a long process. Two important steps are the crystallization of the purified protein and the diffraction of monochromatic X-rays by the protein crystal. The principle of X-ray crystallography is shown in Figure 4. It is the electron clouds of the atoms of the protein that scatter the X-rays. Therefore hydrogen atoms with only one electron are not observed in low resolution X-ray structures. Electrons scatter X-rays in all directions, but maximal positive interference is only observed, if scattered X-rays are in phase with each other (Figure 5). The Bragg equation describes the scattering angles, ϑ, at which maximal interference is observed for a given wavelength, λ, and distance between atoms, *d*.

$$n\lambda = 2d\sin\theta \qquad (6\text{-}1)$$

Many parameters can reveal the quality of an X-ray structure. The resolution of the structure is a very important parameter. In combination with the R and R_{free} it is a good measurement of the quality of the X-ray structure.

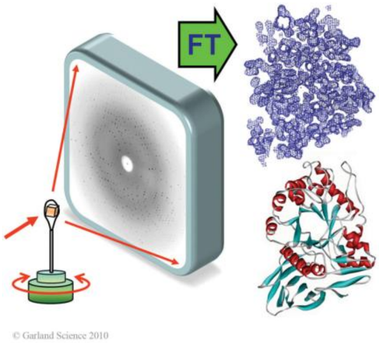


Figure 4 – Overview of the X-ray crystallographic experiment. First a protein crystal diffracts monochromatic X-ray. Second electron densities in real space are calculated from the collected diffraction pattern in reciprocal space. Third atoms are placed within the electron density map. Figure is figure 1-2 from (Rupp 2009).

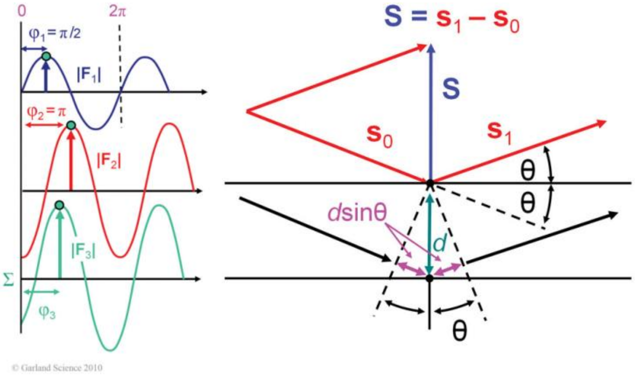


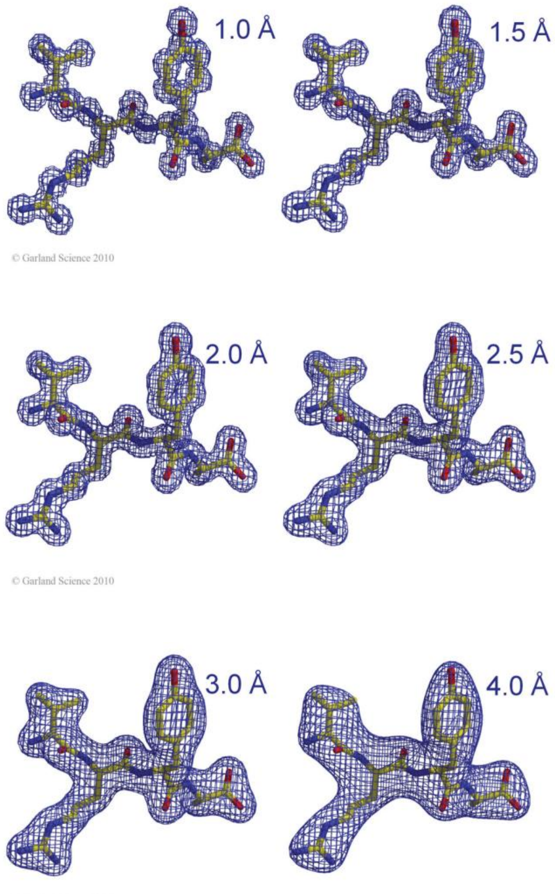
Figure 5 – Interference between monochromatic X-rays with different phases and amplitudes. Maximal positive interference is only observed if waves are in phase. On the right diffraction of X-rays by electrons is shown. If the path difference, *d*sin*ϑ*, due to scattering is a multiple, *n*, of the wavelength, *λ*, then there will be maximal interference of the scattered monochromatic X-rays. The phase difference after scattering equals the double of the distance between the scattering atoms, *d*, multiplied by the sine of the scatter angle, *ϑ*. Therefore the Bragg equation describes the case of maximal positive interference. *nλ* = 2*d*sin*ϑ*. Both figures are from (Rupp 2009). Left figure is figure 6-6 and right figure is figure 6-15.

6.5.1.1.1 Resolution

The resolution of an X-ray structure tells a lot about the quality of that structure. Figure 6 visualizes different resolutions. Atomic resolution refers to a sub 1.2Å resolution. High resolution is 1.2-2.0Å, which is why only sub 1.8Å structures are used for the training set of the side-chain conformation prediction program SCWR and sub 1.7Å structures are used for the underlying backbone-dependent rotamer library(Krivov, Shapovalov et al. 2009). The importance of resolution has also been illustrated for main chain dihedrals by showing the dependence of Ramachandran outliers on resolution(Kleywegt and Jones 1996) and by showing that RMSD and Ramachandran angle differences between NCS related polymer chains increase as the resolution gets worse.(Kleywegt 1996) And finally I have shown myself that there is a correlation between the average B-factor of a structure and the resolution of that structure (Figure 7). I expected this result, as high B-factor atoms gives poor scattering, which in turn gives limited resolution.(Rupp 2009), p. 262) Resolution is independent of space group.(Rupp 2009), p. 235) This justifies later on treating the resolution as a parameter independent of space group. However I have not excluded the effect of oligomeric assembly state, molecular weight and solvent content.

The maximum resolution is given by the smallest distance, d_{min}, at which atoms can be discerned from each. The diffraction limit is determined by the wavelength of the X-ray source, λ, and the maximal diffraction angle $\theta_{\max}$ at which diffraction intensity is still observed. The relationship is given by the Bragg equation.

$$d_{\min} = \lambda / 2\sin\theta_{\max} \qquad (6\text{-}2)$$



© Garland Science 2010

© Garland Science 2010

© Garland Science 2010

Figure 6 – Electron density of a Val-Arg-Tyr-Ala peptide sequence at different levels of resolution. Only at atomic resolution can individual atoms be fully discerned from each other. At resolutions above 2.0Å some side chain dihedral angles are no longer accurately determined. Figure is figure 9-8 from (Rupp 2009).

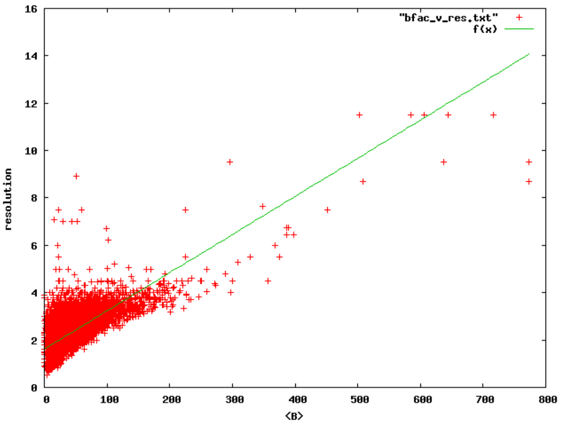


Figure 7 – Correlation plot between resolution and average B-factor of each protein structure in the PDB.

6.5.1.1.2 R and R_{free}

A set of coordinates deposited to the PDB is just an interpretation of electron density, which is calculated from experimental structure factor amplitudes, $F(hkl)$, cf. eq. 6-6. To see if the model fits the experimental diffraction data one can use the R-value.(Rupp 2009), p. 620)

$$R = \frac{\sum_h \sum_k \sum_l \left| F_{obs} - F_{calc} \right|}{\sum_h \sum_k \sum_l F_{obs}} \qquad 6\text{-}3$$

Similar to the RMSD used for comparison of structures the R-value does not reveal, which part of the experimental data set does not fit the model. However the R-value is still important during refinement and upon judging the quality of a structure.

One can always fit a model to experimental data by introducing more parameters. The *F*-test is a common statistical method for evaluating if using more parameters is justified or a case of over fitting.(Farrell, Miranda et al. 2010) In X-ray crystallography the free R-value is used.(Brunger 1992)

$$R_{free} = \frac{\sum_{h \in free} \sum_{k \in free} \sum_{l \in free} \left| F_{obs} - F_{calc} \right|}{\sum_{h \in free} \sum_{k \in free} \sum_{l \in free} F_{obs}} \qquad 6\text{-}4$$

Free refers to part of the dataset (5%) not being used for refinement. If the remaining dataset is over fitted, then the R-value for this dataset (R_{work}) will drop, whereas the R_{free} will stay constant or increase. Over parameterization happens, when disordered solvent molecules and/or multiple side chain conformations are introduced during modeling.

6.5.1.1.3 B-factor

As a measure of the thermal vibration of individual atoms the B-factor is used. (Rupp 2009), eq. 6-17)

$$B_{iso} = 8\pi^2 \left\langle \mu_{iso}^2 \right\rangle \qquad (6\text{-}5)$$

The isotropic B-factor is proportional to the mean square isotropic displacement of each atom from its equilibrium position. Despite isotropic B-factors being inadequate in theory for describing correlated molecular motions (Rupp 2009), p. 644), they often correlate well with normal modes (chapter 9) that themselves correlate well with functional motions; i.e. motions between apo and holo structures, T4L and adenylate kinase structures from different space groups and identical NCS related adenylate kinase chains (data not shown).

Atoms with high B-factors scatter X-rays poorly.(Rupp 2009), p. 262) One would therefore expect structures with high B-factor atoms to have a poor resolution. I have plotted the average B-factor of each protein structure in the PDB against the resolution of that structure (Figure 7). The plot shows a correlation (r = 0.7) between B-factor and resolution.

At atomic resolution anisotropic B-factors are used. If a group of atoms have aligned anisotropic B-factors, then this can be interpreted as the atoms moving in a concerted motion.(Wilson and Brunger 2000) Anisotropic B-factors if available are therefore very helpful for determining domain motions. They are not always available, because anisotopic B-factors require 3 parameters unlike isotropic B-factors, which only require 1 parameter. Anisotropic B-factors therefore cause over parameterization as described in section 6.5.1.1.2 about R and R_{free}.

B-factors reveal the range of motion of atoms of a protein in a crystal. Surface exposed residues will have higher B-factors, whereas buried residues and residues with crystal contacts will have lower B-factors, because their movement is restricted by other nearby residues. However it has been shown that B-factors are not just remnants of the crystal packing; i.e. B-factors of hen and human lysozyme have been shown to correlate despite being crystallized in different space groups.(Artymiuk, Blake et al. 1979)

B-factors are interesting, because they describe thermal vibrations at equilibrium. Although a new method has been introduced for identifying occupancies of distinct side chain conformations(Lang, Ng et al. 2010), an electron density map cannot be used to distinguish a fully occupied atom with a high B-factor from a partially occupied atom with a low B-factor, unless 1) the change between conformations involves just change of side chain dihedral angles, 2) the X-ray structure has been solved at high resolution ($d_{\text{min}} \leq 1.5$Å) and 3) a 0.3-1.0σ electron density level instead of the default 1.0σ level is used. In most cases, even at atomic resolution, are atoms given full occupancy. The CPMG relaxation dispersion experiment can however reveal the size of two distinct conformational populations and thus be a useful experiment to determine, whether an atom has partial occupancy or not. Furthermore the NMR experiment reveals the kinetics of the steady state between the two conformations.

6.5.1.1.4 Molecular replacement

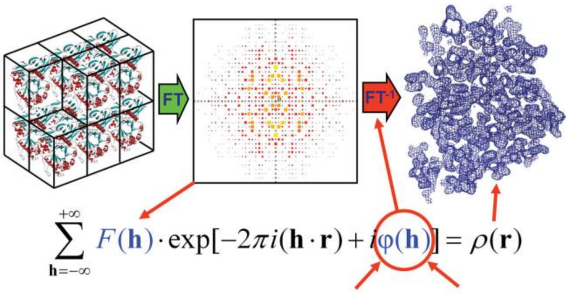
The most common method for solving X-ray structures is molecular replacement (Figure 9). Molecular replacement is a method for solving the phase problem in X-ray crystallography.(Rossmann 1967; Rossmann 2001) The phase problem of crystallography is the problem that the phase of the diffracted X-ray, α_{hkl}, must be known to calculate the electron density as shown in the equation (Rupp 2009), 9-18) below.

$$\rho(x,y,z)=\frac{1}{V}\sum_{h=-\infty}^{\infty}\sum_{k=-\infty}^{\infty}\sum_{l=-\infty}^{\infty}F(hkl)e^{i\alpha_{hkl}}e^{-2\pi i(hx+ky+lz)} \qquad 6\text{-}6$$

ρ is the electron density. *x,y,z* is the fractional coordinate. *V* is the volume of the unit cell. *h*,*k*,*l* is the reciprocal index vector, which is the location of the diffracted X-ray on the detector. *F* is the measured amplitude of the structure factor, which is a summation of the scattering contributions of all atoms in the crystallographic unit cell in the direction defined by the reciprocal index vector. The magnitude of the structure factor is proportional to the square root of the observed intensity. α_{hkl} is the phase of the diffracted X-ray, which is not recorded. As shown in Figure 8 the phases can be calculated from the atomic coordinates of a similar structure. Molecular replacement is the use of phases calculated from a previously determined similar structure. A similar structure could be a sequence identical protein in a different space group or the apo or *wt* form of a holo or mutant structure. Phases are more important than amplitudes in the determination of electron densities and therefore structures determined by molecular replacement suffer from a phase bias (Rupp 2009), meaning that the determination of the position of some atoms in the new structure will be biased by the phases of the structure used for molecular replacement. Today more structures than ever are available in the PDB for use in molecular replacement and MR is the most frequent method for determining phases today (Figure 9). Molecular replacement can be combined with algorithms for protein structure modeling to further strengthen the method and expand the applicability of the method.(Qian, Raman et al. 2007; DiMaio, Terwilliger et al. 2011)

Often the similarity between protein structures can be attributed to the method used for solving the X-ray structure. In this thesis, I hope to clarify whether this is due to molecular replacement or similar experimental procedures. The frequent use of MR is shown in Figure 9. Molecular replacement is a dangerous method to use, if the starting model is of poor quality or plain wrong.(Jones and Kleywegt 2007) MR will fail if the template structure and target structure are different. Usually sequence similarity is a measure of structural difference. But even sequence similar and identical chains can have different conformations, which can be due to small molecule ligands, different space groups and single point mutations. All of these factors can cause large inter-domain motions (e.g. the hinge motion in T4 lysozyme and and adenylate kinase pronounced if different space groups and the hinge motion in calmodulin caused by ligand binding).

If there is a large conformational difference between an apo and a holo structure or a *wt* and a mutant structure, then normal mode analysis can be applied to generate an ensemble of structures. Each of these calculated conformations could be tried as an input structure for molecular replacement, if the apo/wt structure is not suitable itself.(Suhre and Sanejouand 2004) Normal mode analysis can also be used for fitting a high resolution structure into low resolution data.(Tama, Miyashita et al. 2004)


FT
FT⁻¹
$\sum_{\mathbf{h}=-\infty}^{+\infty} F(\mathbf{h}) \cdot \exp[-2\pi i(\mathbf{h}\cdot\mathbf{r}) + i\varphi(\mathbf{h})] = \rho(\mathbf{r})$

© Garland Science 2010

Figure 8 – Illustration of the phase problem in X-ray crystallography. Structure factors can be calculated from atomic coordinates, but electron densities cannot be calculated from structure factors, if the phase is unknown. Figure is figure 9-15 from (Rupp 2009).

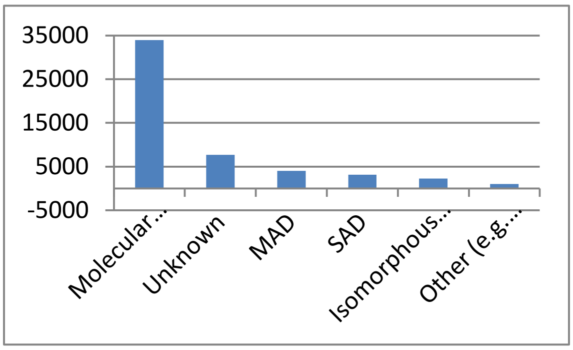


Figure 9 – Frequency of methods for solving the phase problem of X-ray crystallographic determination of protein structure. Drawn with Microsoft Excel.

6.5.1.2 Uncertainty of atomic positions

The uncertainty of atomic positions in X-ray structures has been estimated to be 0.1-0.3Å, 0.5 Å and 0.6-1.0Å by theoretical methods, molecular dynamics and structure comparison respectively.(DePristo, de Bakker et al. 2004) The positions of atoms with high temperature factors are poorly determined from the electron density in particular.(Mowbray, Helgstrand et al. 1999) The RMSD between similar NCS related protein chains has been estimated to be 0.4-0.5Å.(Kleywegt 1996) The RMSD is dependent on the resolution of the structure. Lower resolution causes larger RMSD. Also space group differences rather than uncertainty in the determination of atomic positions cannot be ruled out as a cause of structural differences between NCS related proteins.

Estimates of the uncertainty of atomic position based on molecular dynamics are the least trustworthy, because a measurement of a successful force field is whether the protein will stay folded or not during the simulation. If an MD simulation is allowed to run long enough, then the protein will often unfold or have its secondary structure elements disrupted significantly. Determining differences between structures solved from the same diffraction data is thus a much better estimate of the uncertainty of atomic position. Also it is worth noting that alternate locations are often present in high resolution structures. This structural heterogeneity is evidence that more than one conformation "fits" the experimental data. I have chosen to compare structures solved in the same space group as they are more abundant and readily accessible in the PDB(Berman, Westbrook et al. 2000) than structures refined from the same diffraction data(Joosten, Womack et al. 2009) If solved in the same space group, then the extent of structural variation is not distributed randomly across the protein because of the identical crystal contacts holding the protein in place. Instead it can be located to specific regions of the protein with high conformational freedom; e.g. Gly residues in loops that span across a very large Ramachandran area and residues not restricted in their movement by crystal contacts.

6.5.2 Computational protein structure prediction

I classify protein structure prediction into three categories. One is the prediction of a novel fold from sequence, another is the prediction of a structure for which sequence similar structures already exist and a third is the prediction of a structure differing from an already existing structure by only a single point mutation or the presence of a ligand. The two latter problems are easy, if the goal is to obtain a model close to the target structure, but difficult if the goal is to obtain a model closer to the target structure than the template structure(s).(MacCallum, Hua et al. 2009) The first of the three cases has become less relevant, because the addition of new protein folds to the PDB is not common anymore. Whereas the growth in structures has not stopped, the growth rate of new protein folds(Murzin, Brenner et al. 1995; Orengo, Michie et al. 1997) relative to existing protein folds peaked in 1993-1996 cf. Figure 10 and the absolute growth rate peaked in 2004 with the addition of 146 new folds that year due in large part to the efforts of the structural genomics consortium(Todd, Marsden et al. 2005) and other depositors. With the number of unknown protein folds at an all time low, it should be possible in most cases nowadays to carry out homology modeling from existing sequence similar structures.

Recently protein structure prediction has been carried out by MD simulation(Shaw, Maragakis et al.), but this is a very time consuming method. The dominant method is homology modeling. Biannual critical assessment of protein structure prediction (CASP) has been carried out since 1994. In this forum the Rosetta method developed by David Baker has done particularly well since CASP3 in 1998.(Murzin 1999; Bonneau, Tsai et al. 2001) At CASP6 in 2004 the Baker/Rosetta group submitted a model of a 69 residue protein with an unknown fold and achieved an overall C_α RMSD of 1.6Å.(Bradley, Misura et al. 2005) The assumption of the Rosetta method is that each local protein segment of 3-9 residues of a target structure has a conformation, which is sampled by sequence identical segments in other proteins.(Han and Baker 1995) The overall protein conformation is defined by the conformation of the local segments. By a Monte Carlo procedure the lowest energy conformation of the combined segments as defined by the Rosetta force field is found. The Rosetta force field penalizes poor packing(Bonneau, Tsai et al. 2001); i.e. a low Lennard Jones potential(Jones 1924), many hydrogen bonds and a small solvent accessible surface area is favored(Lazaridis and Karplus 1999). If chemical shifts of residues in a protein have been assigned, then it is possible to reduce the number of conformations that has to be sampled by the Monte Carlo procedure. If the number of Monte Carlo steps can be reduced, then time required for convergence towards an energy minimum will drop. If chemical shifts are known, then this can be accomplished by combing Rosetta with a program, SPARTA, which can predict torsional angles from sequence and chemical shifts(Shen and Bax 2007) into CS-Rosetta. Rosetta carries out pre-filtering of fragments. SPARTA is then used to further exclude fragments with an incorrect conformation before the Monte Carlo procedure. SPARTA is also used to add calculate the chemical shift of the Monte Carlo obtained conformation. The differences between experimental and calculated chemical shifts are then added as a term in the energy function used to identify the lowest energy conformation. By using CS-Rosetta a backbone RMSD as low as 2.03Å can be accomplished for a 147 residue structural genomics target (i.e. a protein with a new fold). And in all cases of protein lengths of less than 100 residues the backbone RMSD falls below 1.0Å.(Shen, Lange et al. 2008) Because knowledge about the chemical shift of a lightly populated conformation can be acquired from the CPMG relaxation dispersion experiment, CS-Rosetta opens up the possibility to deduce the structure of an otherwise sparsely populated conformation.(Korzhnev, Religa et al. 2010)

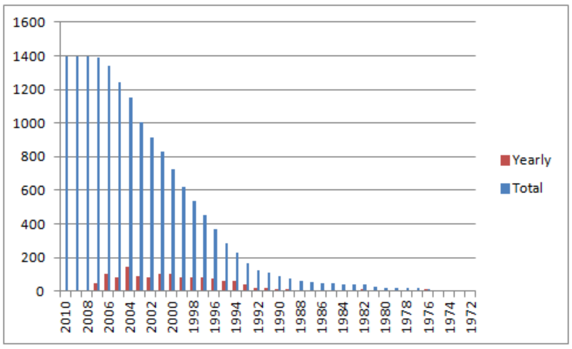


Figure 10 – Normalized totals of SCOP(Murzin, Brenner et al. 1995) folds and structures in the PDB.

6.5.2.1 MD simulations and protein structure prediction

As mentioned earlier, MD simulations have recently been carried out on small proteins to characterize the folding of a protein to its native state.(Shaw, Maragakis et al.) Starting from extended structures the folded structure was obtained on the scale of microseconds using the Amber ff99SB force field described elsewhere in this thesis. A 35 residue alpha helix protein (2f4k) and a 39 residue beta sheet protein (N-terminus of 2f21) folded to their native state (RMSD ≈ 1Å) in 68 and 38μs, respectively. To run a simulation for that long is currently not feasible for larger proteins, which could also fold on a slower ms time scale.

6.6 Protein Structure Comparison

There are several methods in which protein structure similarity can be compared. The most frequently used measurement is the RMSD of the coordinate differences. Here I present the various methods that are used and comment on the strengths and weaknesses of each of them individually. A common weakness of all the methods is that they do not report on the change in flexibility – as measured by changes in B-factors – due to a mutation.(Weinert, Phillips-Piro et al. 2011)

6.6.1 RMSD

A linear regression involves finding the function yielding the smallest sum of squares relative to a set of data points. Likewise super positioning of protein structures involves finding the structural alignment yielding the smallest root mean square deviation (RMSD) between two sets of *N* atom coordinates, *i* and *j*.

$$RMSD = \sqrt{\frac{1}{N}\sum_{i}^{N}\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2 + \left(z_i - z_j\right)^2} \qquad 6\text{-}7$$

Therefore the RMSD is a good measure of structural similarity. Because of inter domain motions the RMSD between large proteins is sometimes larger than that between small proteins. Because of flexible side chains and rigid back bones the C_α RMSD is smaller than the heavy atom RMSD. It is a myth that RMSD is dependent on the size of the protein, because of the size itself.(Cohen and Sternberg 1980) It is true that when comparing random protein structures, the larger proteins will on average have larger RMSDs.(Reva, Finkelstein et al. 1998) However, if homologous proteins are compared, the RMSD will probably only be dependent on size, because larger proteins can display larger conformational changes. If the RMSD was indeed size dependent, then one could however use a size independent measure of protein similarity.(Maiorov and Crippen 1995; Betancourt and Skolnick 2001)

C_α RMSD is the most common method for comparison of protein structures. However, in the CASP competition, there are many measures for structure similarity other than C_α RMSD which are used. The correlation between most of them is strong(MacCallum, Hua et al. 2009). I have therefore decided to limit the number of metrics used for evaluation of protein structure similarity. There is a strong correlation between C_α RMSD and heavy atom RMSD; the latter almost always being larger. Therefore I choose C_α RMSD as my measure of backbone similarity. Prior to taking the root of the mean square deviation the deviations are not scaled by for example the temperature factor of each atom. Thus flexible and rigid parts of a protein will count equally towards the overall RMSD. The problem with RMSD is that an inter-domain perturbation such as a hinge motion or a local perturbation such as a flexible random coil movement can cause large RMSDs despite an otherwise large local and global similarity in the case of the hinge motion and the local perturbation, respectively. If a motif based super positioning method is used, then problems with for example inter domain hinge motions can be overcome.

For my RMSD calculations I use a fast method based on quaternions(Coutsias, Seok et al. 2004) written by David J. Heisterberg at the Ohio Supercomputer Center. All coordinates are weighed equally. Atoms of different secondary structure elements are not treated differently and multi domain proteins are treated like single domain proteins despite the domains being flexible relative to one another. The B-factors of atoms are also ignored despite the fact that high B-factor atoms are more flexible and on average differ the most between two structures.

6.6.2 $\Delta\chi1$

As a measure of sidechain similarity I make use of χ_1 dihedral angles. I prefer this measurement over heavy atom RMSD, because it is less influenced by domain motions than heavy atom RMSD. A disadvantage of looking at side chain dihedrals is that the values are not continuous but rather discrete; they cluster into *trans*, *Gauche*- and *Gauche*+. Another disadvantage of χ_1 values is that they do not report on overall conformational changes caused by a mutation.(Weinert, Phillips-Piro et al. 2011) This is something, which is better gauged by overall RMSD changes.

6.6.3 GDT

The global distance test(Zemla 2003) is the preferred method for structure comparison at CASP. The GDT gives a percentage of residues that are no more distant in two structures than a cutoff value of choice. In the CASP refinement category GDT_TS (GDT total score) rather than GDT is used. GDT_TS is simply the average of GDT using cutoff values of 1,2,4,8 Å.

6.6.4 Multiple Structural Alignment and ensemble RMSD

Superpositioning of an ensemble of conformations and calculation of RMSDs between them is a great measurement of the amplitude/extent of intrinsic dynamics of proteins. Ensembles of proteins can be structurally aligned to each other using multiple structural alignment algorithms. The first of these were developed in the early 1990s.(Vriend and Sander 1991; Orengo, Brown et al. 1992; Shindyalov and Bourne 1998) Instead of comparing RMSDs between two individual structures an “ensemble RMSD” has been suggested instead.(Lindorff-Larsen and Ferkinghoff-Borg 2009) Throughout this thesis I only make use of pair wise structural alignment of coordinates.

6.7 Model proteins

Two proteins appear frequently in this thesis. One is hen egg white lysozyme (HEWL) and the other is bacteriophage T4 lysozyme (T4L) with lengths of 129 and 164 residues respectively. Lysozyme is a glycoside hydrolase. It hydrolyses glycoside bonds of peptidoglycans in the cell wall of bacteria. So unlike penicillin, lysozyme does not prevent the synthesis of peptidoglycans, but rather breaks down the bacterial cell wall. Specifically lysozyme hydrolyses NAG-β-(1,4)-NAG and NAG-β-(1,4)-NAM glycoside bonds - where NAG is N-acetylglucosamine (PDB ID NAG) and NAM is N-acetylmuramic acid (PDB ID MUB) – found in peptidoglycans in the bacterial cell wall. The function of HEWL is to protect from bacterial infection, whereas the function of T4L is to break down the bacterial cell wall at the late stage of the lytic cycle of viral reproduction and permit new viruses to exit the lysed bacterial cell.

6.7.1 Bacteriophage T4 lysozyme

T4 lysozyme was chosen, because its dynamics have been investigated previously.(de Groot, Hayward et al. 1998; Hayward S 1998) The dominant motion between the T4L structures in the PDB from different space groups and between the apo and holo form of T4L is a combination of a hinge closure around the active site protruding the length of the backbone helix running from residues 60 to 80 and a twist of the N-terminal domain (i.e. residues 12-59) relative to the C-terminal domain.(Weaver and Matthews 1987; de Groot, Hayward et al. 1998)

6.7.2 Hen egg white lysozyme

The biological functions of hen egg white lysozyme and T4 lysozyme are identical - i.e. they break down bacterial cell walls and leave bacterial cells vulnerable to osmotic lysis (cytolysis) - but their structures are different. HEWL has a chain length of only 129 amino acid residues and its main catalytic residues are Glu35 and Asp52.(Figure 11) The protonated Glu35 acts as a proton donor, whereas the charged Asp52 acts as a nucleophile to the cleaved glycoside generating a catalytically competent covalent enzyme-glycosyl intermediate.(Vocadlo, Davies et al. 2001) HEWL was the first enzyme to have its structure determined.(Blake, Koenig et al. 1965) HEWL was also one of the first proteins to have its dynamics studied in order to achieve a full understanding of its catalytic function.(McCammon, Gelin et al. 1976) It was also the first protein to which normal mode analysis was applied in order to determine the dominant mode contributing to the apo/holo motion.(Brooks and Karplus 1985) Over the course of 30 years the structure and dynamics of this enzyme has been well studied. Because it is so abundant in the PDB, and because it displays a conformational change upon ligand binding, it is a good choice as a model enzyme.

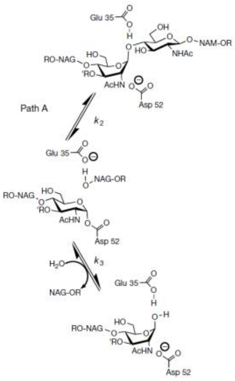


Figure 11 – Catalytic mechanism in HEWL showing functions of Glu35 and Asp52 during glycoside cleavage.(Vocadlo, Davies et al. 2001)

6.8 Parsing PDB and mmCIF files

The structure files in the Protein Data Bank contain more information than what can be found in the x, y and z column of their coordinate section. Often this valuable information is overlooked. Unfortunately providing all useful information is not or has not always been mandatory. Here I present just a few of the most important data items from the header section of PDB and mmCIF files, which I make use of throughout the thesis. It should be noted that parsing information from PDB files can be very difficult and ambiguous when for example residues are missing, whereas the newer mmCIF file format has a clear connection between sequence and structure information.

6.8.1 Title section

The title section contains the PDB record EXPDTA (mmCIF data category exptl), which contains information about experimental technique used for solving a given structure. Unless otherwise stated I strictly make use of structures solved by X-ray diffraction. The PDB record AUTHOR (mmCIF citation_author) contains information about the authors of the publication associated with the structure, if any. mmCIF files also list authors responsible for the data in the deposited structure file (audit_author).

6.8.1.1 PDB REMARK records

The REMARK records of PDB files contain information on such important properties of the structure as resolution (REMARK 2 and 3, mmCIF data categories refine, refine_hist, reflns), refinement software (REMARK 3, mmCIF data item computing.structure_refinement), R-values (REMARK 3), completeness (REMARK 3), temperature and pH (REMARK 200), Matthews coefficient (REMARK 280), crystallographic symmetry (REMARK 290), transformation matrices to get from asymmetric units to biological units (REMARK 350), missing and zero occupancy residues and atoms (REMARKs 465, 470, 475, 480).

6.8.2 Crystallographic and Coordinate Transformation Section

The crystallographic section contains the important PDB record CRYST1, which contains information on the parameters of the unit cell such as side lengths, angles, space group and Z value (the number of polymer chains in one unit cell) for calculating the Matthews coefficient, if it is not given.

6.8.3 Primary Structure Section

The primary structure section contains sequence database references to UniProt (PDB record DBREF, mmCIF data item struct_ref.db_name), which is the gold standard of protein sequence databases.(Apweiler, Bairoch et al. 2004; The UniProt 2011) The database cross referencing allows for the identification of residues mutated relative to the wild type sequence, which is provided in the SEQADV PDB records / struct_ref_seq_dif mmCIF data category.

6.8.4 Connectivity Section

The connectivity section (and connectivity annotation section) contains the all important information on the type of interaction between the observed atoms (PDB records CONECT and SSBOND and mmCIF data item struct_conn). The position of an atom alone does not always reveal the character of a bond between two atoms.

6.8.5 Coordinate Section

The most important part of a structure file is of course the section containing the coordinates of the atoms that make up the structure. However, more attributes are associated with each atom than just a coordinate in 3D space. Each atom also has an associated occupancy, which is a useful measure of the population sizes of two conformations. All too often this value is set to 1, because individual conformations are not distinguished. Zero occupancy residues unfortunately is not a common phenomenon despite the occasional contradiction between a full occupancy atom and observation of its electron density. Each atom is also associated with a B-factor (introduced in sections 6.4.2.2 and 6.5.1.1.3). The B-factor can be either isotropic or anisotropic. For various reasons B-factors between two structures are not directly comparable. One of them being that refinement with REFMAC prior to version 5.5.0042 only includes the TLS contribution to the isotropic B-factor.(Winn, Ballard et al. ; Collaborative 1994; Murshudov, Vagin et al. 1997; Winn, Isupov et al. 2001; Vagin, Steiner et al. 2004)

7 Chapter 2 - HEWL and T4L structural variability

7.1 Introduction

The field of protein design and engineering is of ever-increasing importance for the development of novel therapeutics and industrial biocatalysts. As a result, theoretical methods for analyzing and rationalizing the effect of amino acid substitutions on protein characteristics are in high demand. Often, the structures of mutant proteins are solved as part of protein engineering or protein design projects; however, the effects of the vast majority of point mutations are analyzed solely with the help of *in silico* models of mutant proteins. Such models are increasingly being used with structure-based energy calculation algorithms to obtain predictions of the effect of the point mutation on the binding characteristics, stability(Gilis and Rooman 2000; Guerois, Nielsen et al. 2002; Johnston, Søndergaard et al. 2011), protein-protein association rates(Selzer, Albeck et al. 2000; Kortemme and Baker 2002), and pK_a values(Tynan-Connolly and Nielsen 2006; Tynan-Connolly and Nielsen 2007). It is therefore of interest to study the structural effect of point mutations to ensure that the models used with structure-based energy calculation algorithms are accurate and that one arrives at a correct prediction for the right reason when attempting to correlate experimental measurements with predictions.

A central question in this area is related to the volume of the protein structure that is affected by the introduction of a point mutation. For reasons of convenience (and for lack of a tested and proven alternative), it is often assumed that the structural changes associated with a point mutation are limited to the mutated residues; i.e. one only adjusts the positions of inserted/changed atoms in order to arrive at a model of the mutant protein (Krivov, Shapovalov et al. 2009). In the case of a mutation to a smaller residue this approach may work; however the assumption breaks down when larger residues are inserted in the core of a protein, since some atoms must move ‘out of the way’ to accommodate the newly inserted atoms. Structural rearrangements due to the insertion of larger residues can be modeled using rotamer-optimization algorithms based on Monte Carlo simulated annealing, graph theory or dynamic dead-end elimination algorithms(Desmet, Maeyer et al. 1992). These methods are implemented in several software packages such as Rosetta(Das and Baker 2008) and ORBIT(Ross, Sarisky et al. 2001) and Yasara(Krieger, Joo et al. 2009). A couple of models that allow for the optimization of the backbone conformation have been constructed; however, the benchmarking of such algorithms has, to my knowledge, been relatively limited.

The problems associated with modeling point mutations in proteins are well appreciated in the field of protein engineering and design. Nevertheless, to my knowledge, there has been no systematic effort to characterize and catalogue the effects of point mutations on protein structures and thus provide a dataset for benchmarking protein-mutation/protein-design algorithms. Here, I present a detailed analysis of the effects of point mutations on the structures of Hen Egg White Lysozyme (HEWL) and bacteriophage T4 Lysozyme (T4L) protein structures in the Protein Data Bank (PDB). I examine the extent of structural change observed for the mutations and compare these changes with the changes that are induced by other factors affecting the crystal structure (space group, temperature, pH).

7.1.1 Measuring the effect of protein point mutations

The effect of point mutations can be observed experimentally. Point mutations can influence thermal stability(Bava, Gromiha et al. 2004), catalytic rates (O'Meara, Nielsen in preparation), pK_a values(Tynan-Connolly and Nielsen 2007), binding constants(Schreiber and Fersht 1995) and folding rates(Chiti, Taddei et al. 1999; Ladurner and Fersht 1999). When electrostatics properties are studied and pK_a values manipulated, then mutations causing charge reversal, neutralization or shift between a polar and apolar environment can be carried out(Tynan-Connolly and Nielsen 2006). Cavities can be introduced by mutating a large residue to a smaller one(Eriksson, Baase et al. 1992).

7.1.2 Modelling point mutations

Typically models of point mutations are constructed using backbone-specific rotamer libraries(Chinea, Padron et al. 1995; Simon, Word et al. 2000; Krivov, Shapovalov et al. 2009), and these are sometimes optimized using either local or global energy minimization or full molecular dynamics analyses. Alternatively, point mutations can be modeled using full protein design packages that allow for the optimization of the conformations of nearby residues using a rotamer sampling scheme.(Smith and Kortemme 2008)

7.1.3 Choice of HEWL and T4L

HEWL and T4L were chosen for detailed analysis, because structures of them are abundant in the PDB. Among 346,059 structure pairs I have identified in the PDB more than 10,000 of those can be attributed to HEWL alone. My dataset includes 392 HEWL and 172 T4L structures. Many structures of *wt* HEWL were solved at different physiochemical conditions, while many mutant structures of T4L were solved at somewhat similar physiochemical conditions. With more than 400 structures solved, the HIV-1 protease is one of the most abundant proteins in the PDB. However, I did not use HIV-1 protease for analysis as the biological unit consists of two chains which complicate the analysis. A mutation in one chain will inevitably have an effect on the neighboring chain and the presence of two chains gives rise to variation not only caused by intrinsic chain dynamics, but also by the movement of chains relative to each other. For the same reason hemoglobin, with its four chains in the biological unit, has been excluded from this analysis. The major histocompatibility complex is another abundant protein in the PDB. It was not used for this analysis, as it is often bound to other proteins. Taking into account protein-protein interactions complicates the analysis too much. Overall, T4L and HEWL are great proteins for this analysis, as they are very abundant in the PDB and are monomers in solution, and the asymmetric unit of most space groups is also constituted by a single chain. The two proteins supplement each other well for my analysis; many mutant structures of T4L are available, whereas many *wt* structures of HEWL are present. T4L allows the study of the effect of mutations, whereas HEWL allows the study of the influence of space group, pH, and temperature and so on. The full datasets are available at [www.proteinkemi.dk/thesis/hewl.txt] and [www.proteinkemi.dk/thesis/t4l.txt]. The following abbreviations are used in the online tables: *T* = temperature, res = resolution, SG = space group, SM = starting model (not traced back), V_m = Matthews coefficient, Z = number of HEWL chains in unit cell

7.2 Materials and methods

7.2.1 Super positioning of protein structures

Super positioning of protein structures involves finding the best overlap of the coordinates of two (or more) structures in space. If all coordinates are weighed equally, then the RMSD of the coordinate differences is the lowest achievable RMSD. The super positioning of structures is here based on the quaternion method (code by David J. Heisterberg from The Ohio Supercomputer Center, 1990, unpublished results) (section 6.6.1). When I calculate the heavy atom RMSD and compare a *wt* structure to a mutant structure, I do not use the side chain atoms of the mutated residue for the super positioning.

When performing heavy atom structural alignment, it is assumed that hydrogen bond network optimization has been carried out to determine the orientation of Asp, Glu, His, Asn and Gln. Symmetrical atoms ($C_{\gamma1}$/$C_{\gamma2}$ of Val; $C_{\delta1}$/$C_{\delta2}$ of Leu; $C_{\delta1}$/$C_{\delta2}$ and $C_{\varepsilon1}$/$C_{\varepsilon2}$ of Phe and Tyr) are included in the heavy atom structural alignment, even though they are easily mistaken for each other due to a 180 degree rotation of a single of their side chain dihedral angles. In fact, if the χ_1 rotation is fast, then the two pairs of side chain atoms in Phe and Tyr are indistinguishable even by NMR. Only in low resolution structures are the terminal side chain atoms of Val and Leu mistaken for each other by a misinterpretation of the χ_1 and χ_2 angles by 180 degrees.

7.2.2 Calculation of RMSD and $\Delta\chi_1$ values

The super positioning is only carried out once. Upon calculation of the RMSD for a subset of the structure the super positioning for the whole structure is used.

The same set of atoms (C_α or heavy atoms) is used for super positioning and calculation of RMSD. Both calculations are either based on C_α atoms or heavy atoms.

During a conformational change, the χ1 angles of side chains and backbone atom positions will change. The C_α RMSD is a good gauge of backbone conformational changes, while the changes in χ1 angles provide a good estimate of side chain flexibility. In addition, if there is no concerted motion in the protein, then heavy atom RMSD is also a good measure of side chain flexibility. But T4L and HEWL both undergo hinge motions that are large scale concerted conformational changes of the position of two domains relative to each other. Using heavy atom RMSD would be a measure of this hinge motion rather than side chain variation.

Plotting C_α RMSD and average change in χ_1 angles against each other reveals what structures are most similar; both in terms of backbone and side chain variation.

7.2.3 Selection of structures, atoms and residues

Only HEWL structures solved by X-ray diffraction are compared. The presence of other polymer entities (e.g. antibodies) excludes the structure from analysis; non polymer entities (e.g. ions) and sugar molecules are allowed to be present, unless they are covalently bound to HEWL (i.e. 2b5z, 1h6m, 1uc0). The presence of a crosslinker in 2htx and 2hu1 excludes these two structures from analysis. Structures with modified non standard residues are also excluded from the data set (i.e. succinimide in 1at5, isoaspartate in 1at6, methylated lysines in 132a, carboxymethylated cysteine in 1rcm). In addition, a few structures are excluded from the analysis due to their poor resolution (1lzh (6Å), 2lzh (6Å), 1bhz (3.9Å)). The great variation in side chain conformation between the low resolution (1lzt) and atomic resolution structure of HEWL (3lzt) may be due to the fact that 1lzt was derived from 6Å resolution data and the errors in the parent model were thus inherited.(J. M. Hodsdon 1990) As a result I exclude 1lzt from the analysis. 1aki has been solved by molecular replacement using the 6Å resolution structure 2lzh as a starting model. Therefore, I also exclude 1aki from analysis.

T4L structures with covalently attached spin labels are excluded from analysis. This includes 2nth (Leu118 spin label), which has a perturbed F helix (residues 109-113) due to the presence of the spin label.

High temperature factor atoms are used for super positioning and calculation of RMSD, but zero occupancy atoms are not used (e.g. side chain atoms in loop 99-103 in HEWL structure 2f30). The position and conformation of zero occupancy atoms are merely modeled and not based on experimental observation. Due to this fact they are treated as unobserved atoms. The same rule applies to the calculation of χ1 dihedral angles. In the case of alternative conformers with equal occupancies, the first conformer in the coordinate section of the structure file is used.

The loop between residues 67-73 is a source of large variation between HEWL structures (Figure 13). For this reason the residues in loops 45-49, 67-73, 99-103 with high temperature factors are in some analyses treated as a separate group of flexible residues (e.g. Figure 17 in section 7.3.2.2).

7.2.4 Identification of dehydrated structures

Low solvent has a significant impact on the unit cell dimensions(Bernal, Fankuchen et al. 1938) and thus the effective crystal contacts. Structures solved at low humidity are identified by plotting the resolution dependent Matthews factor against the resolution.(Kantardjieff and Rupp 2003). From the plot (Figure 12), dehydrated structures can easily be identified.

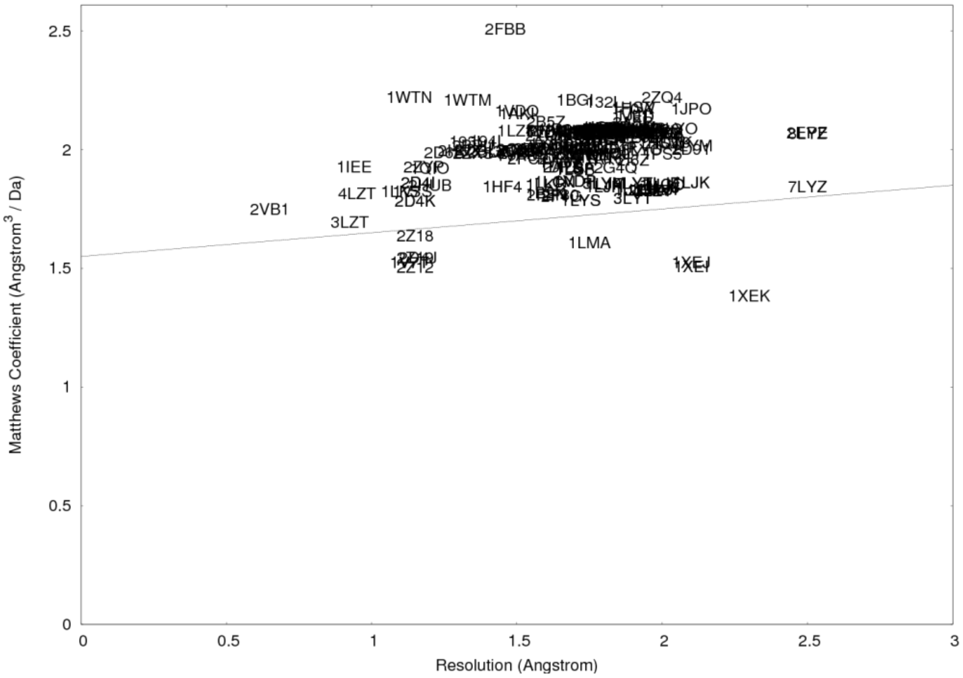


Figure 12 - The Matthews coefficient of each HEWL structure is plotted as a function of the resolution of that structure. Structures with low solvent content are those with a low Matthews coefficient.

7.3 Results

A better understanding of mutation-induced protein structure changes is of high importance in the continuing efforts to understand and improve protein design and engineering, for assessing the impact of Single Nucleotide Polymorphisms (SNPs) on protein structure and function and for the general understanding of protein structure plasticity. In the following parts of this result section I examine the effects of mutations in HEWL based on a comparison of 25 single point mutants and 190 wild type HEWL structures from the PDB. First, I assess the structural variability of *wt* HEWL structures as a function of space group, pH and temperature. Next, I compare the differences between *wt* and mutant HEWL structures, paying particular attention to the extent of structural rearrangement caused by the mutation of a residue. Finally, I examine the influence of ligand binding on the structure of HEWL and discuss the implications of my findings for the modeling of point mutations in X-ray structures and for the use of X-ray structures of mutant proteins in protein engineering projects.

7.3.1 Properties of residues contributing to structural variance

The variation for each of the 129 residues in the HEWL structures in the dataset is shown in Figure 13. One could argue that residues with a high flexibility should weigh less upon structure comparison; I have however refrained from doing this. Figure 14 shows that the average difference in atom position between HEWL structures is dependent on the average B-factor of the two atoms being compared to each other. Nevertheless, I have chosen to let all residues weigh equally or split the data into a set of flexible and a set of rigid residues (e.g. Figure 17).

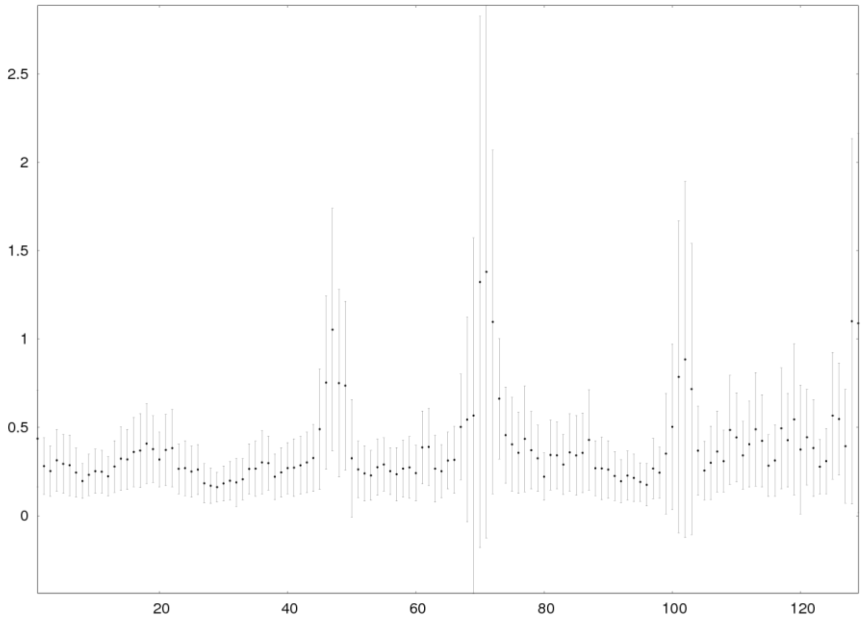


Figure 13 - Residue fluctuations among HEWL *wt* structures. On the x-axis are residue numbers. On the y-axis is the C_α RMSD. The points show the average RMSD per residue by comparison of all *wt* structures. The vertical lines are standard deviations.

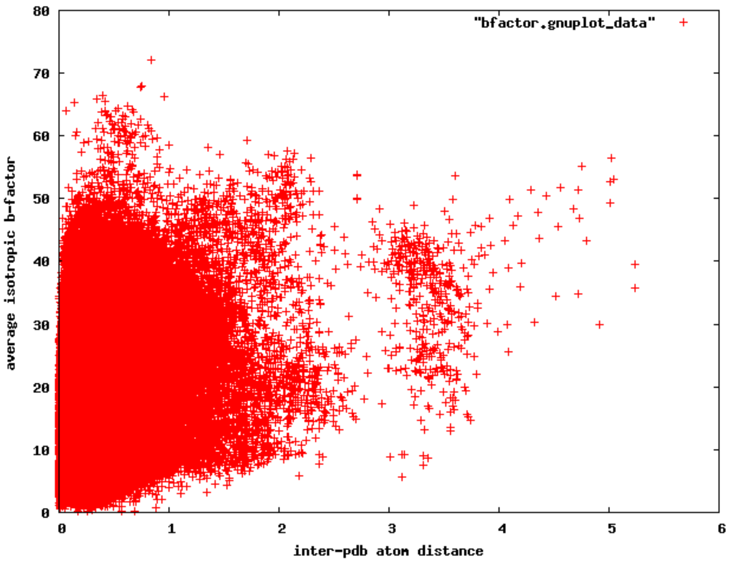


Figure 14 - Scatter plot of the distance between identical atoms in all combinations of two aligned HEWL structures plotted against the average isotropic temperature factor of the two atoms.

7.3.2 Assessing the variability in HEWL *wt* protein structures

When measuring the effect of a point mutation on a protein structure, it is essential to first observe the variability of the wild type structure to define a proper baseline for the subsequent comparisons with mutant structures. In this respect HEWL presents an ideal data set, since 254 structures of *wt* HEWL are present in the PDB. In the following I report on the comparison of HEWL *wt* structures.

The average Cα RMSD and average heavy atom RMSD for all pair wise comparisons for *wt* HEWL structures is 0.51Å (n = 17955, std dev = 0.27) and 1.15Å (n = 17955, std dev = 0.30), respectively.

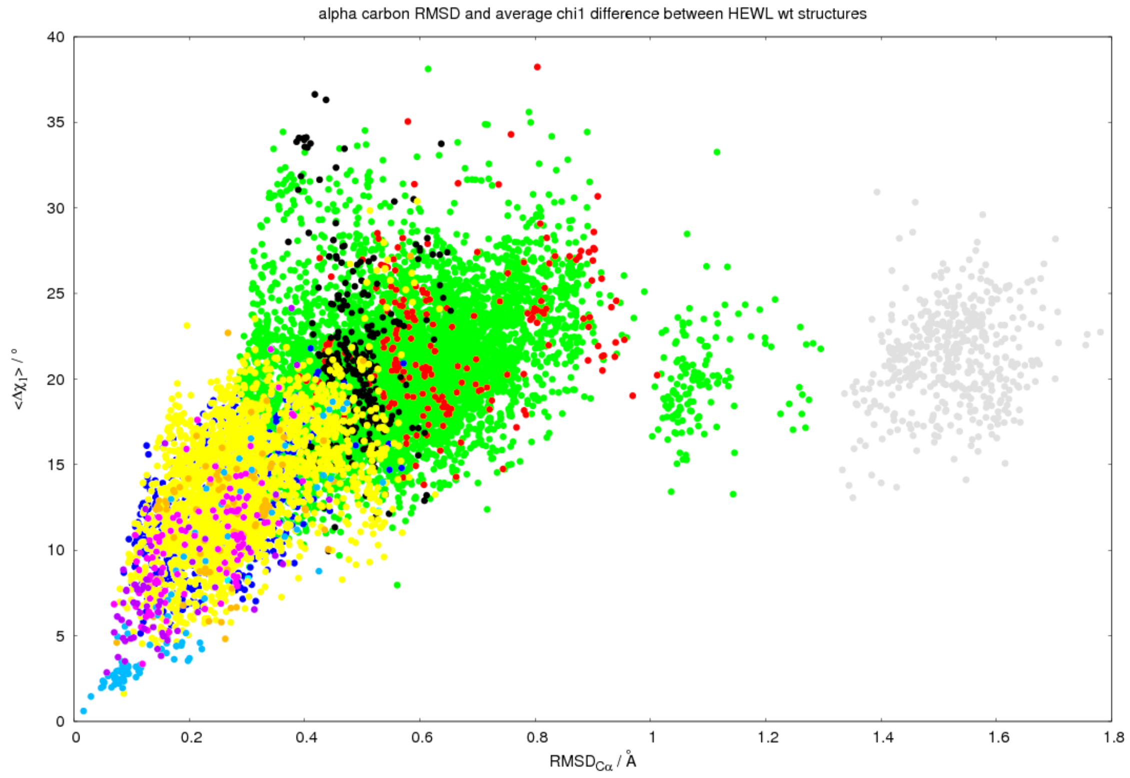


Figure 15 – C_α RMSD vs $<\Delta\chi_1>$ for all HEWL wild types in the PDB excluding exceptions mentioned in the text. The results are summarized in Table 1. •Grey: The presence of thiocyanate or iodide in the 67-73 region in HEWL induce a significantly different conformation of HEWL. One of the structures in this set crystallizes in the P_{21} space group. Furthermore, only one of the HEWL chains in the asymmetric unit occupies a different conformation, whereas the other chain occupies a 'normal' conformation. •Green: Comparisons of two structures from different space groups. •Blue: Both structures in the pair crystallize in the P_{21} space group. •Cyan: both structures in the pair crystallize in the same space group, but the authors of the structures are different and the starting models used for refinement/molecular replace are different. •Yellow: both structures in the pair crystallize in the same space group, but the authors are different.

Colour	Ions	S.G.	Deh.	Diff. CC	Author	Model	Ref.	n	$RMSD_{C\alpha}$ (Å)	$RMSD_{heavy}$ (Å)	$\Delta\chi_1$ (°)
Grey	Yes	-	No	No	-	-	-	453	1.52 +/- 0.08	1.70 +/- 0.09	21.36 +/- 2.83
Green	-	Different	-	-	-	-	-	6926	0.55 +/- 0.16	0.96 +/- 0.17	20.12 +/- 3.66
Red	-	same (P_{21},P_1)	Yes	-	-	-	-	206	0.64 +/- 0.15	1.05 +/- 0.16	22.14 +/- 4.09
	-	same (P_{21},P_1)	No	Yes	-	-	-	304	0.49 +/- 0.05	0.98 +/- 0.08	20.96 +/- 4.75
Dark blue	-	Same	No	No	diff	diff	-	814	0.25 +/- 0.08	0.56 +/- 0.10	12.32 +/- 2.87
Cyan	-	Same	No	No	same	diff	-	10	0.30 +/- 0.12	0.62 +/- 0.13	11.10 +/- 2.72
Yellow	-	Same	No	No	diff	unkn	-	2700	0.28 +/- 0.09	0.61 +/- 0.13	13.35 +/- 3.46
Light blue	-	Same	No	No	same	unkn	same	103	0.27 +/- 0.08	0.54 +/- 0.15	12.32 +/- 3.44
Orange	-	Same	No	No	same	unkn	Diff	97	0.17 +/- 0.12	0.32 +/- 0.24	6.38 +/- 5.00
Pink	-	Same	No	No	diff	same	-	112	0.22 +/- 0.08	0.46 +/- 0.13	10.51 +/- 3.17
Violet	-	Same	No	No	same	same	-	56	0.16 +/- 0.09	0.40 +/- 0.17	7.94 +/- 3.89

Table 1. Averages and standard deviations of RMSD and $\Delta\chi_1$ differences in HEWL. The data points are shown in Figure 15. •S.G. = space group, •Deh. = dehydrated, •Diff. CC = same spacegroup, but different crystal contacts due to non-equivalent chains, •Ref. = primary reference.

7.3.2.1 Space group and crystal contacts

No property causes a larger structural variance than the space group. Figure 15 shows the correlation between the C_α RMSD and the $\Delta\chi_1$ when comparing two different HEWL chains from the same or different X-ray structures. The green points of Figure 15 represent comparisons between structures from different space groups. The large structural differences are explained by interactions with different crystal contacts. Since different crystal contacts cause the large RMSD it is not surprising to find that non-equivalent chains in the asymmetric unit (ASU) of the monoclinic and triclinic structures of HEWL have just as large RMSDs (blue points) as the chains from different space groups (green points). Having established that different crystal contacts is the single most important factor contributing to structural differences, my analysis is reduced to a comparison of chains from identical space groups with isotropic crystal contacts.

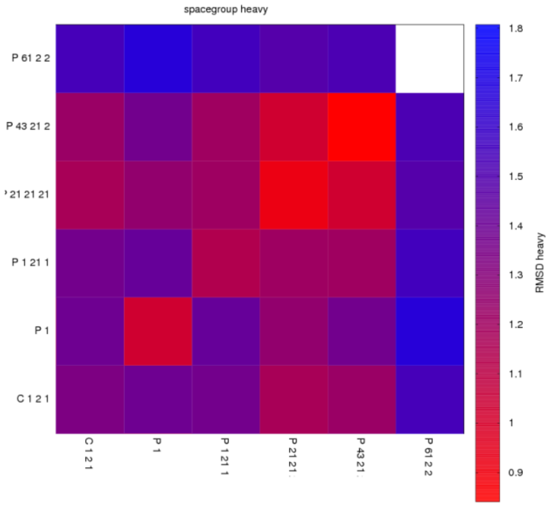


Figure 16 - Each square shows the average heavy atom RMSD of alignments between all structures from the space group in the row and column. The lowest average RMSD is observed, when structures from the P 43 21 2 space group are compared against each other. A white square is used if only one single structure has been solved within the space group.

7.3.2.2 Author and starting model

It is common for labs to consistently use the same starting model, when solving a new X-ray structure of HEWL by molecular replacement. For cases in which the publishers of two structures are different and in which the starting model is unknown for either of the two structures, the cases are treated distinctively. Those cases are colored with yellow in Figure 15.

Structures with identical crystal contacts that are solved by the same lab or solved by molecular replacement from the same starting structures (Figure 15 pink points, Figure 18) generally have a lower RMSD than structures solved in different labs and from different starting models (Figure 15 cyan points, Figure 18). A few exceptions are colored with red points. These red points are comparisons with structures solved at low solvent (1v7t, 1xei, 1xej, 1xek, 2z12, 2z18, 2z19, 2d4j, 1lma). Dehydration (low solvent) causes deformation of the residues in the 67-73 loop, although it is not as pronounced as in 1b2k:A, 1lcn:B, 1lkr:B; e.g. Arg73 does not have an inverted ϕ angle (Figure 25).

Interestingly, the heavy atom RMSD is almost equal to the C_α RMSD in many of the cases where the structures are solved by the same lab or from the same starting model (Figure 15 pink points). This may show that a) side chain conformations are highly dependent on the starting model, b) the similarity could be an artifact of the same lab using identical buffers and other experimental settings or c) it could illustrate that backbone and side chain dihedral angles of high temperature factor residues in non-atomic resolution structures are always modeled with the same conformation within each lab.

The reason I do not include symmetrical atoms of for example Tyr in heavy atom RMSD calculations is that often one lab will name symmetrical atoms in one way, while another lab has the isotropic atoms flipped. This would give rise to a smaller heavy atom RMSD, when comparing structures from the same lab, and a larger heavy atom RMSD, when comparing structures from different labs.

If the conformation of an X-ray structure solved by molecular replacement is dependent on the starting model, then the RMSD between a starting model and its derived models should be lower than the RMSD between the same starting model and random HEWL *wt* structures. This is illustrated in Figure 17. The dataset is split up between flexible and fixed residues, because the experimental determination of the position of the flexible residues is associated with a larger margin of error. I wanted to make sure that calculated differences were not due to a less prevalent conformation being selected by the X-ray crystallographer solving the structure. Preferably in the case of two conformations being sampled, the X-ray crystallographer should identify both as valid conformations(Fraser, Clarkson et al. 2009) with different occupancies.

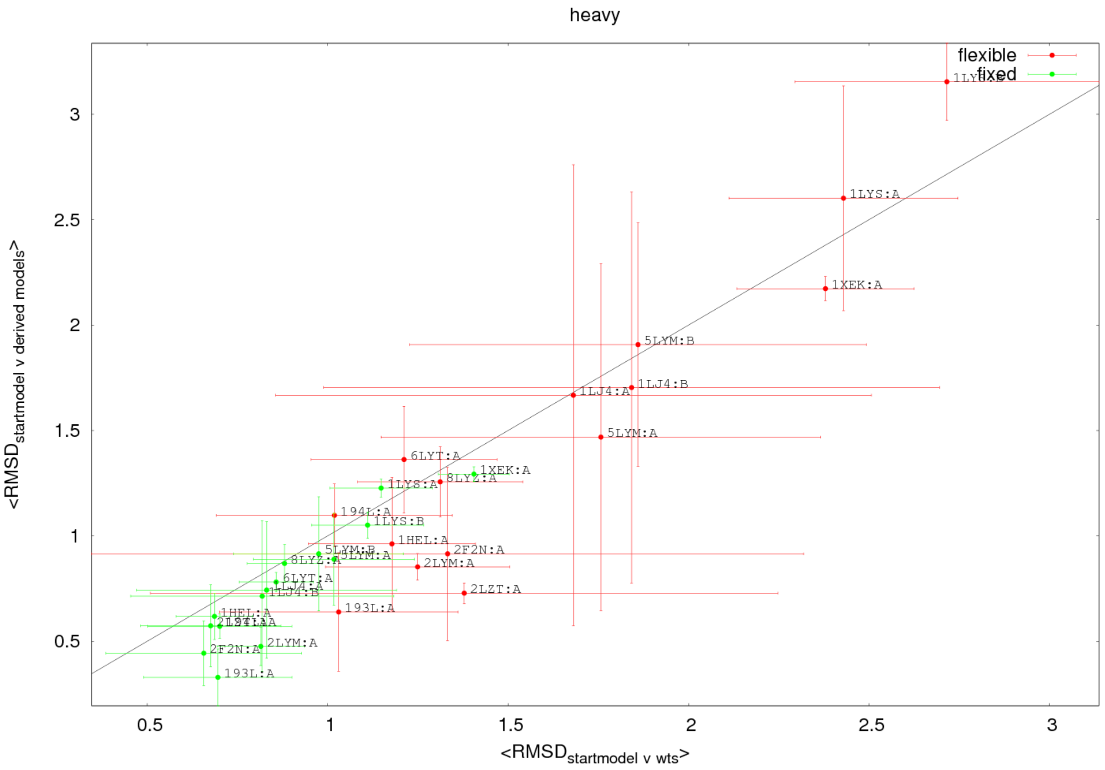


Figure 17 – Comparison of HEWL structures – used as starting models when solving other HEWL X-ray structures by molecular replacement – to their derived structures and other *wt* structures not derived from them. The dataset is split into fixed (Chargaff, Lipshitz et al.) and flexible (red) residues. On the abscissa is the RMSD between the start model and all other wild types. On the ordinate is the RMSD between a start model and structures derived from it. If the derived models are more similar to their starting model than other wild types, then all points should lie below the diagonal. This is a consistent result for the buried residues (Chargaff, Lipshitz et al.), whereas the solvent exposed residues (red) have higher B-factors and seem to have slightly more independently determined conformations.

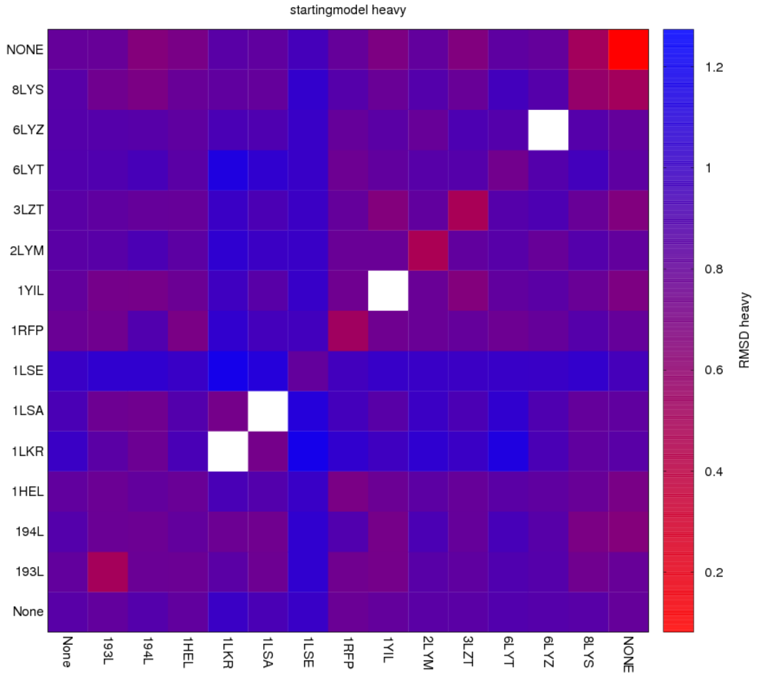


Figure 18 - Each square shows the average heavy atom RMSD between P 43 21 2 HEWL structures with the starting models given in the row and column. If only one structure is based on a starting model, then no average RMSD is calculated and the diagonal squares are colored in white.

	different				same				difference (different-same)		
	n	RMSDCα (Å)	RMSDheavy (Å)	Δχ1 (°)	n	RMSDCα (Å)	RMSDheavy (Å)	Δχ1 (°)	RMSDCα (Å)	RMSDheavy (Å)	Δχ1 (°)
Author	3626	0.27	0.59	13.03	266	0.222	0.43	9.18	0.06	0.16	3.85
SM	824	0.25	0.56	12.30	168	0.20	0.44	9.65	0.05	0.12	2.65

Table 2 - Effect of the author and starting model on HEWL *wt* structures. If the author or starting model is different, then all three properties (alpha RMSD, heavy atom RMSD and χ_1 difference) are higher than in the case in which the author and starting model are similar. The C_α RMSD is almost equally dependent on the author than the starting model, whereas the side chain positioning (heavy atom RMSD and χ_1 difference) is slightly more dependent on the starting model than the author.

7.3.2.3 Effect of resolution and physiochemical properties

Next I focus on the effect of resolution and miscellaneous physiochemical properties on the given conformation of HEWL. Considering that structural variance is mostly explained by crystal contact differences, all other possible effects are only examined within the P 4_3 2_1 2 space group. The P 4_3 2_1 2 space group was chosen because a) most of the available HEWL structures are from this space group, b) the asymmetric unit (ASU) only contains one chain and c) the structures within this space group are very similar on average (Figure 16).

7.3.2.4 pH differences

Hydrogen bond patterns and thus side chain conformations are expected to be dependent on pH. As a result, a correlation between RMSD and the difference in pH between two crystals may be observed. Figure 19 shows a poor positive correlation (*r* = 0.20) between RMSD and the pH difference. The largest pH difference is less than 4 pH units and the difference in pH between most structures is less than 1 pH unit. Because all the data points are clustered at low pH differences, one cannot expect to see a large correlation. Figure 20 shows the distribution of pH differences, which is not normal distributed.

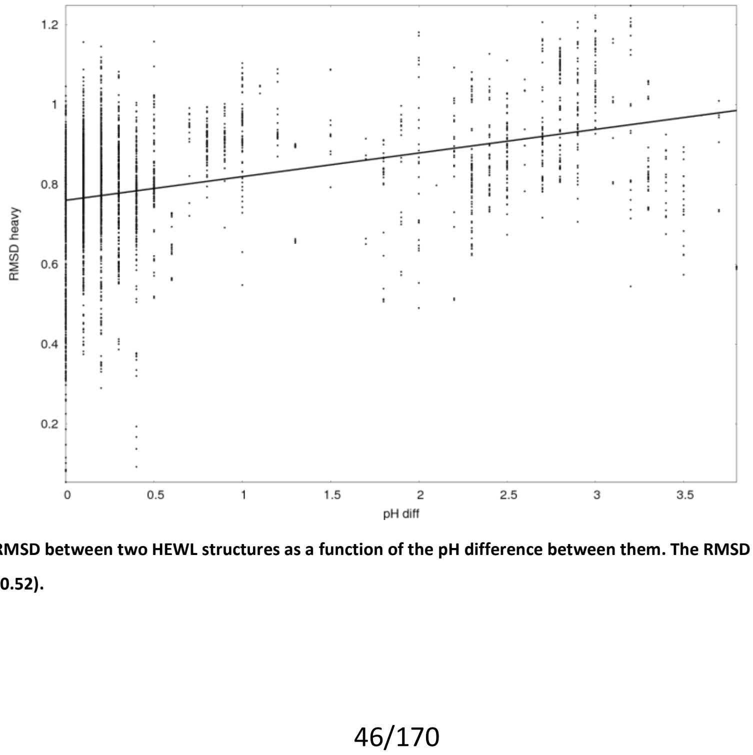


Figure 19 - Heavy atom RMSD between two HEWL structures as a function of the pH difference between them. The RMSD increases as the pH difference increases (*r* = 0.52).

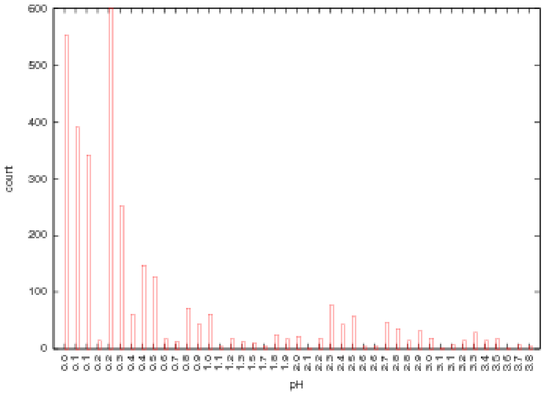


Figure 20 – Distribution of pH differences between HEWL structures.

7.3.2.5 Temperature differences

The distribution of temperature differences between structures is even more skewed than that of pH differences because structures are either cryogenic or solved at room temperature. Figure 21 shows a plot of temperature difference against RMSD. A positive correlation (*r* = 0.36) is observed, but it is not supported by intermediate temperature differences.

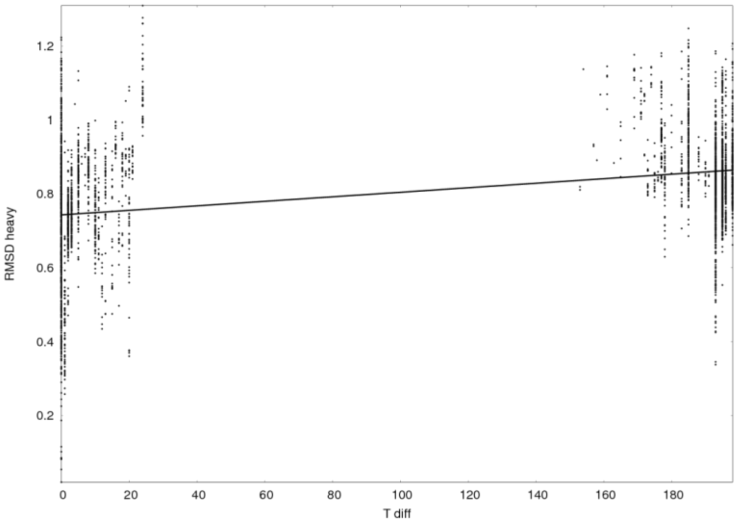


Figure 21 - Heavy atom RMSD between two structures as a function of the temperature difference between them. The RMSD increases as the temperature difference increases (*r* = 0.36).

7.3.2.6 Solvent content

The Matthews coefficient is a measure of the solvent content in the crystal. It is the volume of the unit cell divided by the weight of the polymers in that unit cell. Thus a high Matthews coefficient is the equivalent of a low solvent content. If a crystal has a low solvent content and is more tightly packed, it is not unlikely that the solvent exposed side chains will be restricted in their range of motions. Therefore I expect conformational differences between structures that have significantly different Matthews coefficients. This is exactly what I observe as shown in Figure 22. In the case of crystal structures from different space groups, the structural differences might also be attributed to entirely different crystal contacts, whereas the same trend when comparing structures from the same space group cannot be disputed.

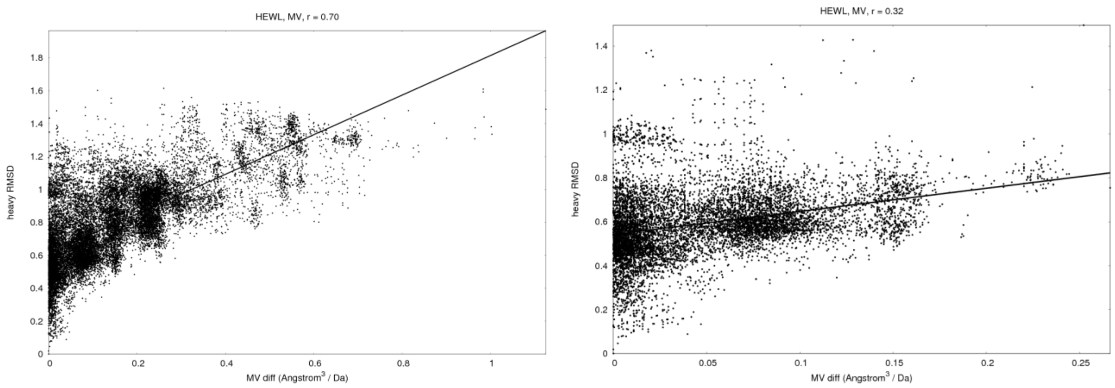


Figure 22 - Heavy atom RMSD between two HEWL structures as a function of the difference in Matthews coefficient between these two structures. Left: Between structures from all space groups (left). Right: Between structures from the same space group. As the Matthews coefficient increases, so does the RMSD. The correlation between the heavy atom RMSD and Matthews coefficient difference between structures from all space groups (left) and identical space groups (right) is *r* = 0.70 and *r* = 0.32, respectively.

7.3.3 Identifying mutant-induced structural variability in HEWL

The HEWL dataset is very sparse in terms of mutants. The only starting models from which at least one wild type structure and two mutant structures have been derived are shown in Table 3. Therefore it is difficult to determine the structural effects of mutations in HEWL if any.

starting model	*wt*	Mutants
2lzh (low res, P 21 21 21)	1aki (P 21 21 21)	1heq,1heo,1hep,1hen,1hem,1her (P 43 21 2)
1rfp (P 43 21 2)	1uih (P 43 21 2)	1fn5,1flq,1ior,1flw,1fly,1ioq,1flu,1iot,1ios (P 43 21 2)
6lyz (P 43 21 2)	1azf (P 43 21 2)	1lzd,1lze,1lzg (P 43 21 2)

Table 3 – HEWL mutants and the starting model used to solve their structure by molecular replacement.

2lzh is a low resolution structure and in the case of 1rfp and 6lyz only a single *wt* structure has been solved using one of these structures as a starting model. It is therefore not possible to carry out a large scale comparison of mutants and wild types derived from the same starting model.

All of the mutant structures in this study – with the exception of 1lza – have been solved by the research groups of Imoto or Matthews (Table 4). The *wt* structure most similar to the mutants is in all cases 1rfp, 1uig, 1vdt or 1hel.

mutant	author of mutant	mutation	wt structures with CA RMSD below 0.1	wt structure most similar to mutant	CA RMSD	starting model of mutant	starting model of most similar *wt* structure
1kxw	Imoto	N27D		1uig	0.118	1rfp	1hel
1kxy	Imoto	D18N	1rfp,1uig	1uig	0.086	1rfp	1hel
1uic	Imoto	H15A		1rfp	0.113	1rfp	1hel
1uid	Imoto	H15F		1uig	0.154	1rfp	1hel
1uie	Imoto	H15G		1rfp	0.114	1rfp	1hel
1uif	Imoto	H15V		1rfp	0.130	1rfp	1hel
1flq	Imoto	G117A	1rfp,1uig	1rfp	0.067	1rfp	1hel
1flu	Imoto	G67A		1uig	0.144	1rfp	1hel
1flw	Imoto	G71A		1uig	0.192	1rfp	1hel
1fly	Imoto	G102A		1vdt	0.138	1rfp	unknown
1fn5	Imoto	G49A		1uig	0.160	1rfp	1hel
1ios	Imoto	M12F		1rfp	0.107	1rfp	1hel
1iot	Imoto	M12L	1rfp,1uig	1rfp	0.095	1rfp	1hel
1ir7	Imoto	I78M	1rfp,1uig	1uig	0.094	1rfp	1hel
1ir8	Imoto	I58M		1rfp	0.153	1rfp	1hel
1ir9	Imoto	I98M	1rfp	1rfp	0.099	1rfp	1hel
1hem	Matthews	S91T		1hel	0.111	1hel	1hel
1heo	Matthews	I55V	1hel	1hel	0.087	1hel	1hel
1her	Matthews	T40S	1hel	1hel	0.093	1hel	1hel
1lzd	Kumagai	W62Y	1lza	1lza	0.088	6lyz	6lyz

Table 4 – Selected HEWL structures from Imoto and solved by molecular replacement using 1hel or a derivative of 1hel as a starting model.

And for all of these *wt* structures the starting model is 1hel. The mutant structure is therefore heavily dependent on the starting structure used in molecular replacement. Otherwise the mutants would have been more similar to any random *wt* structure. Therefore it does not make sense to compare a mutant with anything else but the starting structure (or a structure generated from the same starting model as the mutant).

Even when comparing mutant structure with their starting model, one cannot be certain that all the structural variation can be designated as originating from the mutation. This is illustrated strikingly by comparisons between wild type structures where one structure has been used as the starting model for molecular replacement (Figure 17).

Here I try to answer if single point mutations have a significant effect on the structure of HEWL. Figure 23 shows that mutant structures are more identical to each other and selected *wt* structures than *wt* structures in general despite the structures being from the same space group (P 4_3 2_1 2). This is evidence of the importance the starting model has on the final reported conformation. Figure 24 shows the C_α RMSD between HEWL structures from the same space group (P 4_3 2_1 2) in the absence of the outliers discussed previously (1b2k:A, 1lcn:B, 1lkr:B). The figure shows that the wild type structures are more different from each other than the mutant structures are from the wild type structures. This can perhaps be attributed to what starting model was used when building the mutant structures by molecular replacement. The conclusion is that experimental methods for solving the structure give rise to a larger apparent structural variation than single point mutations. Either the structural consequence of single point mutations is small and/or the conformations available for HEWL to sample and for the X-ray crystallographer to choose among many and very different. Both would explain why apparently *wt* structures are more different from each other than single point mutants are from each other and selected *wt* structures.

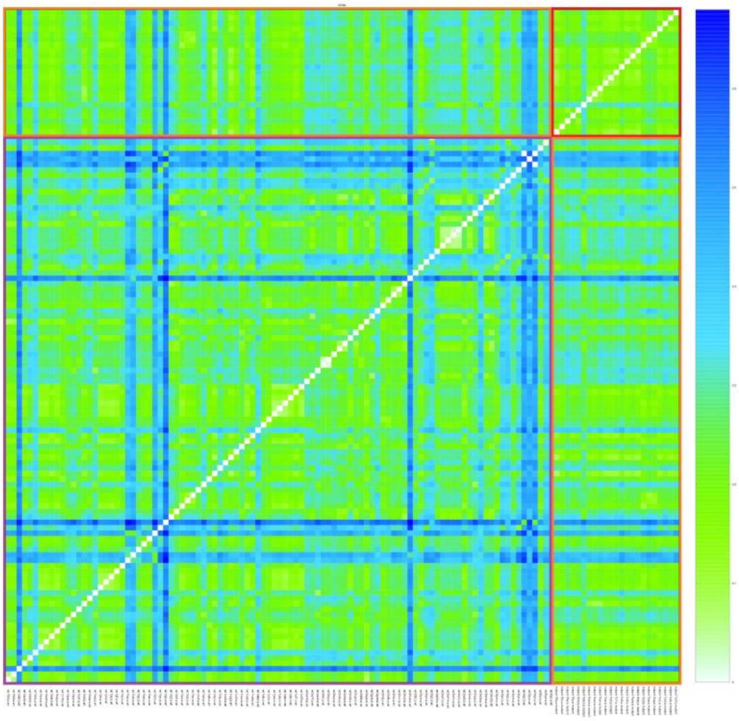


Figure 23 – C_α RMSDs between HEWL structures in the P 4_3 2_1 2 space group in the absence of ligands. Each square represents an RMSD between two HEWL structures. The matrix is symmetrical around the diagonal. The diagonal is colored in white. White is the color used to represent an RMSD of zero. The comparison of wild types against wild types is shown in the bottom left corner (violet square). The comparison of mutants against mutants is shown in the top right corner (red square). The comparison of wild types against mutants are shown in the orange rectangles.

Mutations do not cause larger structural effects in either of the two HEWL domains than what is seen for the wild types. In (Figure 24) a comparison of the average C_α RMSD between, on the one hand, an ensemble of *wt* structures, and on the other hand, a randomly selected *wt* structure (black) or a mutant (colored), reveals the structural variation in each of the two domains to be identical and larger on average for the *wt* structures.

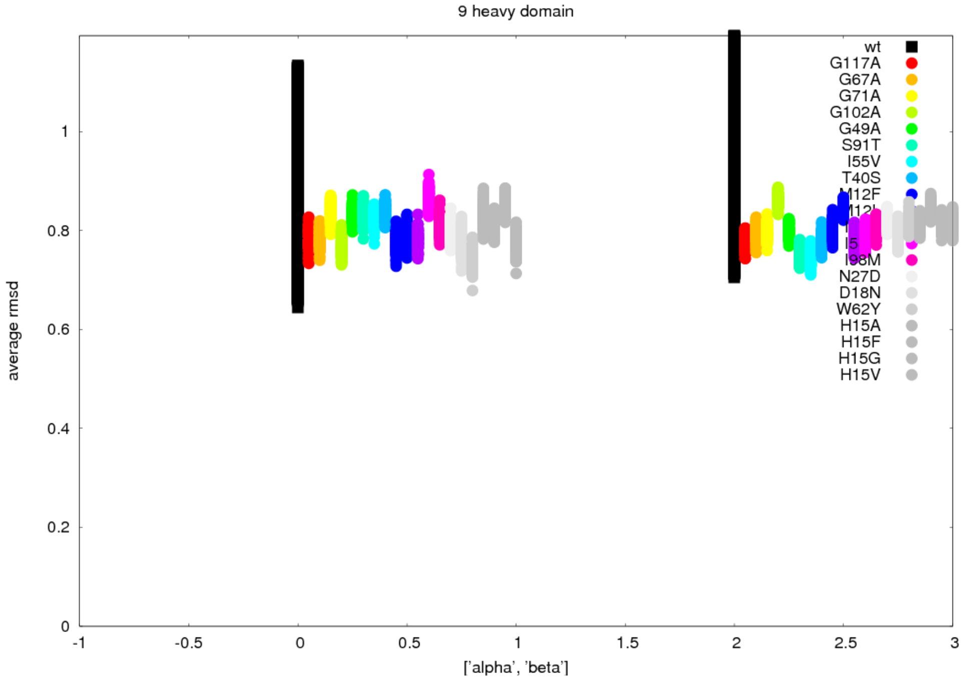


Figure 24 - The average RMSD of a sample of *wt* structures compared against *wt* structures (black) and mutants (colors) are calculated. The RMSD of alpha and beta domain residues are shown on the left and right of the plot respectively.

7.3.4 Unexpected effects of ligands on the HEWL structure

7.3.4.1 Monovalent ions in the P 2_1 space group

Here I look at the effect small monovalent ions have on the structure of HEWL. As previously discussed in section 7.3.2, there are a few chains that give rise to very large RMSDs, when compared to other HEWL structures (orange points in Figure 15). These chains are 1b2k:A, 1lcn:B and 1lkr:B. All three chains are from the P 1 2_1 1 space group and they all have a neighboring chain in their ASU. Other proteins crystallized in the monoclinic (P 1 2_1 1) and triclinic (P 1) space group also have two chains in the ASU with non-identical crystal contacts. For these structures the RMSD between chains from the same ASU is higher than the RMSD between equivalent chains from different X-ray structures. However, the three chains mentioned above are significantly different from any other chains in the monoclinic space group. This is mostly due to a variation in the loop between residues 67 and 73 (Figure 25). These residues have high temperature factors, and furthermore three of them are Glycine or Proline residues (Gly67, Pro70 and Gly71). Glycine residues are spread out wide across the Ramachandran landscape (Figure 1). The three chains all have large monovalent ions at the loop residues; thiocyanate in 1lkr and iodide ion in 1b2k and 1lkr. Iodide does not have the same effect on structures solved in the P 4_3 2_1 2 space group (1gwd, 1vat, 2d91). HEWL has been crystallized with the smaller bromide ion (vdW radius of 1.82Å relative to the 2.06Å vdW radius of the iodide ion), but unfortunately it does not crystallize in the tetragonal and not the monoclinic form (Dauter and Dauter 1999), 1azf(1998)), and therefore it is unknown whether the size of the ion has a structural effect or not. 2d4k is solved in the P 2_1 space group in the presence of chloride ions (vdW radius of 1.67Å). Residues 67-73 of chains 2d4k:A and 2d4k:N in the ASU have slightly different conformations, which is attributed in particular to the φ/ψ angles of Pro70 and Gly71, but the effect is not as pronounced as with the iodide ion. Unlike the iodide ion, the binding of chloride at residues 67-73 is symmetric; meaning that the ion is found at the same position in the two chains in the ASU.

It cannot be readily explained why iodide and thiocyanate trap HEWL in an exclusive conformation (1b2k and 1lkr). The different loop conformations have previously been explained by an interaction between Gly71 O and the iodide atom(Vaney, Broutin et al. 2001). Vaney et.al. suggest that some anion binding sites are space group independent and only dependent on the nature of side chains, whereas others are space group dependent. In the case of the asymmetric binding of iodide, the binding site in the vicinity of Gly71 appears to be space group dependent.

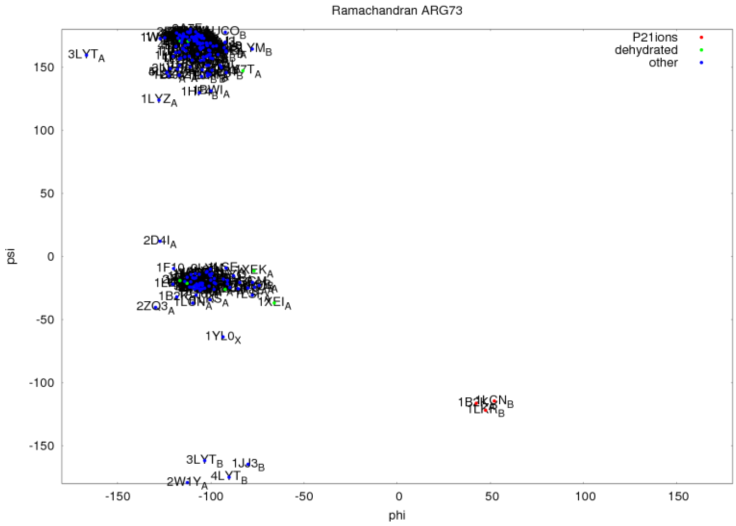


Figure 25 - Ramachandran plot of Arginine 73. 1b2k:A, 1lkr:B, 1lcn:B from the P 2_1 space group to which monovalent ions are bound have an inverted ɸ angle.

7.3.4.2 Urea and NAG

Unexpectedly urea has no effect on HEWL as can be seen from 2f30 solved in the triclinic space group. Whether the substrate NAG has a significant effect on the conformation of HEWL remains unanswered, because the HEWL structure in the presence of NAG has only been solved by powder diffraction (1ja7) or as the E35Q mutant (1h6m,2war).

7.3.5 Assessing the variability in T4L *wt* and mutant structures

As is the case with HEWL, it is the space group that dictates the conformation of T4L (Figure 26). The largest RMSDs are observed when the space groups are different (Chargaff, Lipshitz et al.). However, in some cases, the protein-protein association in the unit cell is similar across space groups, which gives rise to small RMSDs.

Most of the T4L structures in the PDB are – unlike the HEWL structures – mutant structures. Some of the T4L structures have been engineered to dimerize via an intermolecular disulfide bond. In other cases residues 42 and 121 have been mutated to cysteine to bridge the hinge gap. Comparisons of structures with intermolecular disulfide bonds or the CYS42-CYS121-disulfide bond are shown in light grey (Figure 26).

In some cases, chains from the same unit cell will have very different crystal contacts. These comparisons of chains within a unit cell and non-equivalent chains across unit cells from the same space group are colored in black (Figure 26).

Once more it is observed that the conformations are very much dependent on author and starting model. The cases in which the author or the starting models are the same have much lower RMSDs than the average pair wise comparison of other T4L structures (Table 6).

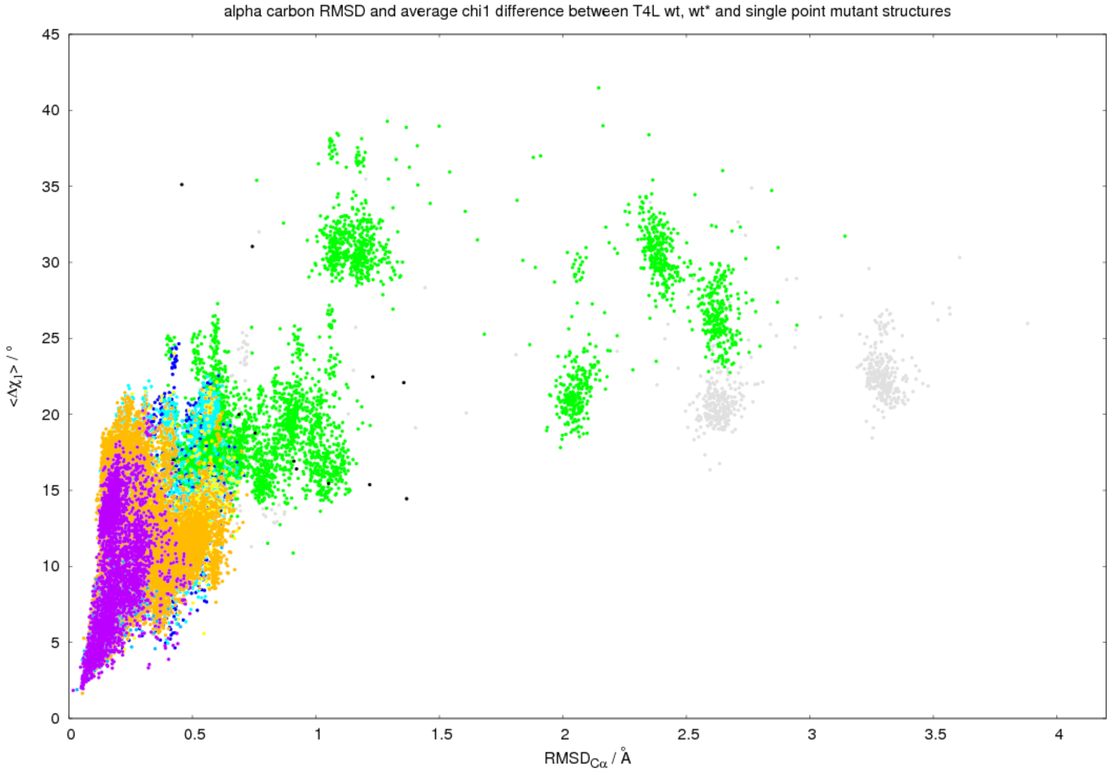


Figure 26 - Cα RMSD vs. heavy atom RMSD (LEFT) and Cα RMSD vs. <$\Delta\chi_1$> (RIGHT) for all T4L *wt*, *wt and single point mutant structures in the PDB excluding exceptions mentioned in the text. The coloring is similar to that of HEWL except for grey. Grey represents structures with intermolecular or intramolecular disulfide bonds (i.e. the I3C mutant).**

Colour	disulf	SG	CC	Author	Model	Ref.	n	RMSD$_{C\alpha}$ (Å)	RMSD$_{heavy}$ (Å)	$\Delta\chi_1$ (°)
Grey	Yes	-	-	-	-	-	866	2.19 +/- 1.12	2.43 +/- 1.03	20.51 +/- 3.48
Green	-	different	-	-	-	-	4991	1.01 +/- 0.66	1.28 +/- 0.61	21.14 +/- 5.58
⬛	-	same	Yes	-	-	-	18	0.78 +/- 0.34	1.05 +/- 0.28	18.20 +/- 6.19
Dark blue	-	same	No	diff	diff	-	1230	0.40 +/- 0.12	0.65 +/- 0.13	13.23 +/- 3.89
Cyan	-	same	No	same	diff	-	6104	0.32 +/- 0.13	0.56 +/- 0.12	12.74 +/- 3.48
Yellow	-	same	No	diff	unkn	-	1631	0.39 +/- 0.13	0.66 +/- 0.12	13.09 +/- 2.78
Light blue	-	same	No	same	unkn	same	24123	0.24 +/- 0.10	0.45 +/- 0.13	10.20 +/- 3.15
Orange	-	same	No	same	unkn	diff	452	0.17 +/- 0.07	0.34 +/- 0.12	6.66 +/- 1.84
Pink	-	same	No	diff	same	-	9	0.30 +/- 0.07	0.58 +/- 0.12	11.45 +/- 1.06
Violet	-	same	No	same	same	-	3038	0.19 +/- 0.07	0.38 +/- 0.15	9.91 +/- 3.83

Table 5 – The table shows average RMSDs for selected subsets of T4L structures. SG = space group, CC = same spacegroup, but different crystal contacts due to non-equivalent chains, ref. = primary reference

	Different				same				difference (different-same)		
	n	RMSDCα	RMSDheavy	$\Delta\chi1$	n	RMSDCα	RMSDheavy	$\Delta\chi1$ (°)	RMSDCα	RMSDheavy	$\Delta\chi1$
Author	5425	0.41	0.68	13.48	47881	0.26	0.49	10.99	0.15	0.19	2.49
SM	11476	0.34	0.61	13.17	3666	0.22	0.41	10.28	0.12	0.21	2.89

Table 6 – The table summarizes the effect of the authors and starting models on T4L *wt*, *wt and single point mutant structures. If the authors or starting models are different, then all three properties (C_α RMSD, heavy atom RMSD and average χ_1 difference) are higher than in the case in which they are similar. The C_α RMSD is slightly more dependent on the author than the starting model, whereas the side chain positioning (heavy atom RMSD and average χ_1 difference) is slightly more dependent on the starting model than the author. The RMSDs are in Å. The angles are in degrees.**

In a few cases the large coordinate RMSDs between T4L structures can be explained by the hinge motion in T4L. In most other cases, the structural differences are not due to domain movements, but rather structural changes in each domain. In order to determine, whether the structural differences between T4L mutant and *wt* structures are attributable to changes in the large or the small domain and the presence of a mutation in each of the two domains, I plotted the measures of structural difference in each domain (Cα RMSD, heavy atom RMSD, $|\Delta\chi_1|$) against each other on separate axes (Figure 27). Because structures are more similar to their starting model than other structures, I only use the starting models for comparison by structural alignment.

Structures with mutations in the small and large domain are shown in green and red respectively. The measure of structural difference is shown on the y-axis and the x-axis for the small and large domain respectively. Therefore I would expect green points to lie above the diagonal and red points to lie below the diagonal, if the mutation has the largest effect on the domain, in which it is located. This is indeed the general pattern I observe for all three parameters (C_α RMSD, heavy atom RMSD, $|\Delta\chi_1|$) in Figure 27. However the many exceptions also show that structural variation between wild types and mutants cannot always be explained by mutational effects. Therefore one should be skeptical of conclusions in the literature reached about structural changes caused by mutations.

The inner grey diagonal lines in the three plots (Figure 27) are lines of linear regression through all of the data points. The outer lines are the 95% confidence intervals of these linear regressions. In all three plots it is observed that the RMSD is larger in the small domain than in the large domain. This is especially pronounced in the last of the three plots (Figure 27c, χ_1 variation). It is a very common result of NMA calculations that the smaller of two domains have longer vectors. T4L is no exception. Therefore it is no surprise that the largest variation is observed in the small domain. This adds further to the evidence that the structural variation observed is due to spontaneous conformational sampling rather than being caused by single point mutations.

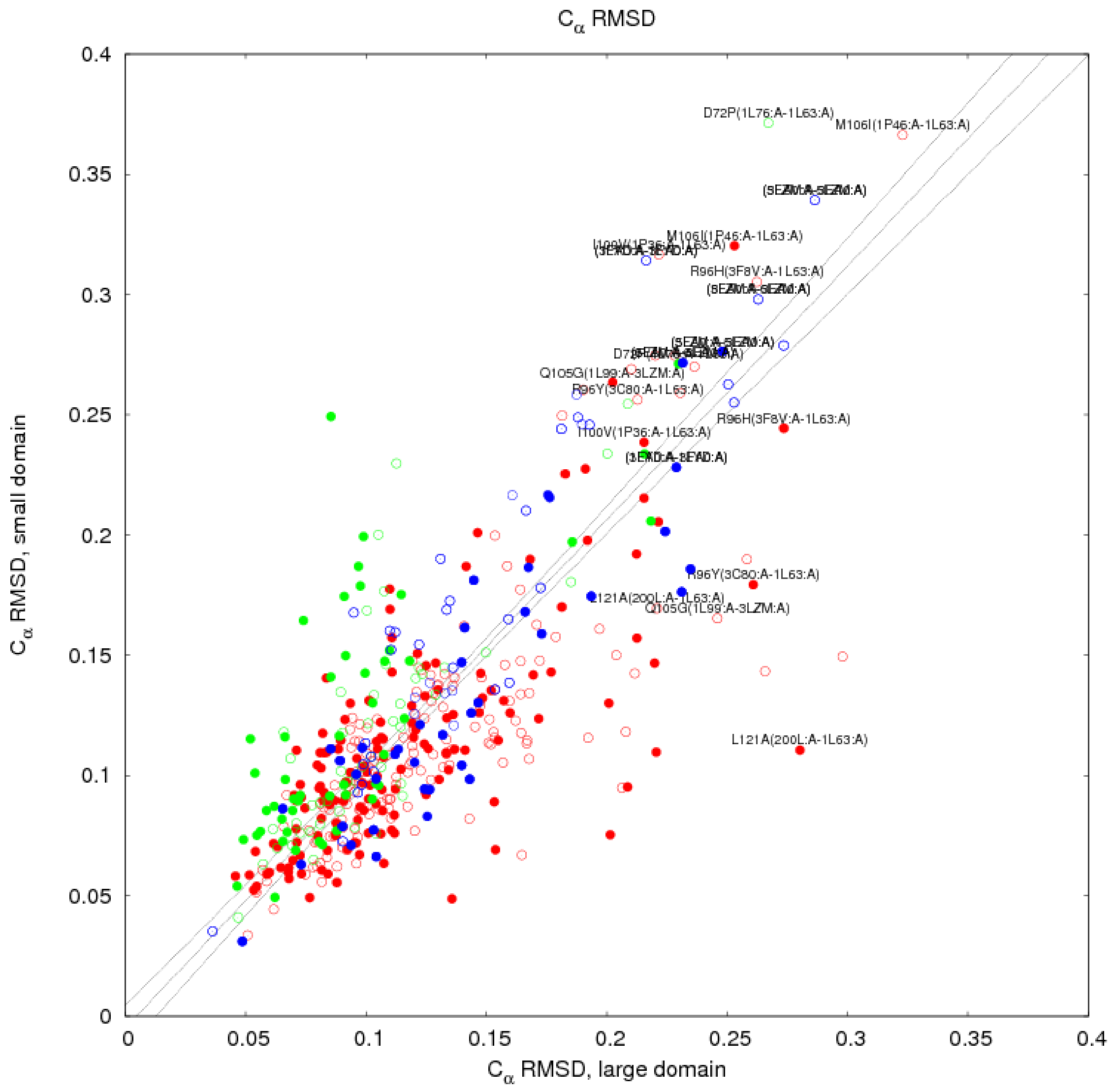


Figure 27a – Figure caption text is on the next page.

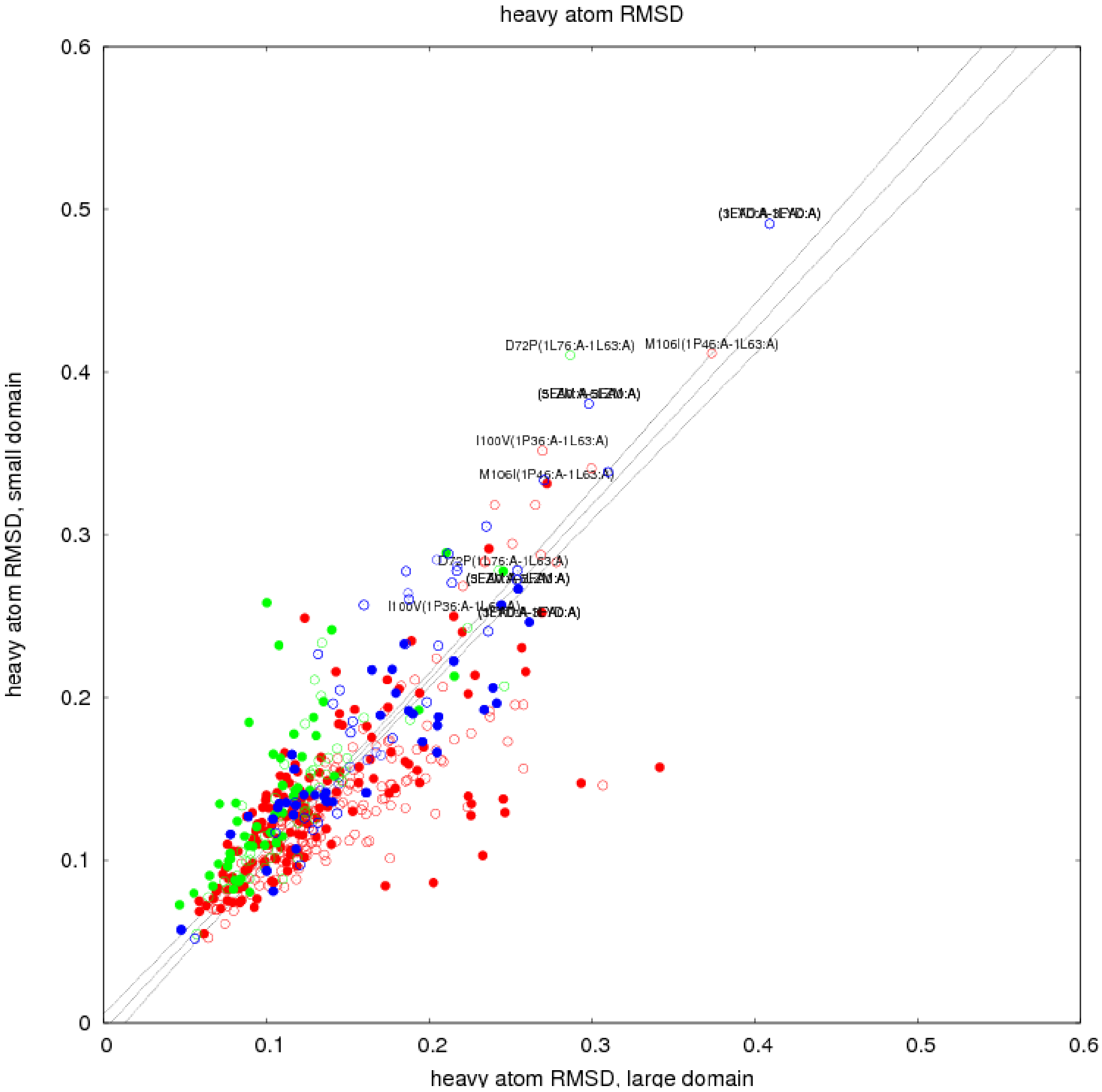


Figure 27b – Figure caption text is on the next page.

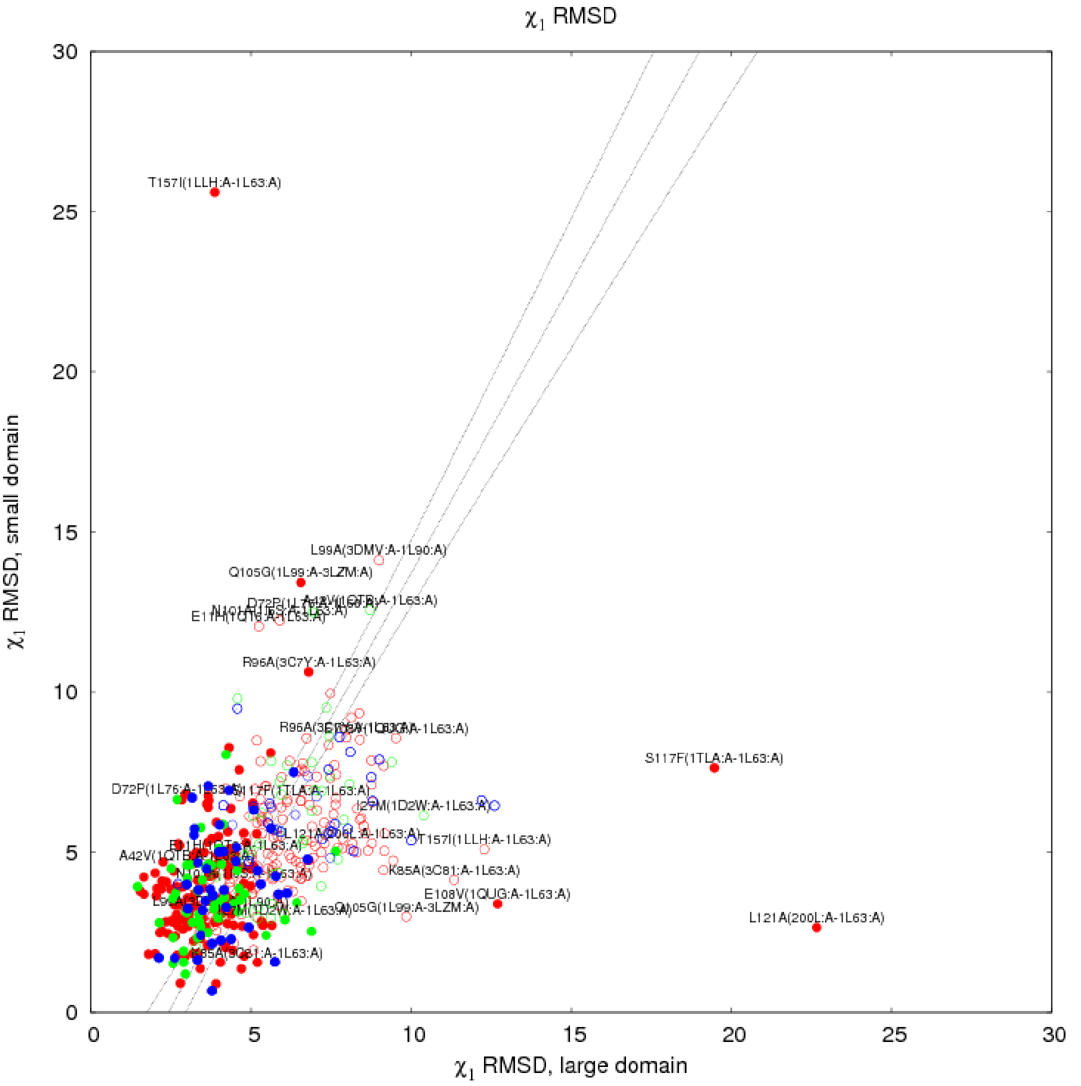


Figure 27c - Scatter plot of the RMSD of the large domain (x-axis) and the RMSD of the small domain (y-axis). The RMSD is shown for the C_α coordinates (top), the heavy atom coordinates (middle) and the chi1 angles (bottom). The coordinate RMSDs of large, small, buried and exposed residues are calculated from structural alignments to all heavy atoms. The structural alignment is performed between each mutant and its start model if known. If the start model is not known, then the comparison is made between 3lzm and 1l63 in the case of *wt* and wt* derived structures, respectively. Red and green circles represent mutations in the large and small domain, respectively. Filled and empty circles represent RMSDs of buried and exposed residues, respectively. Mutants are compared to their molecular replacement starting model. If the starting model is not given, then Cysteine-free mutants (*wt mutants) are compared with 1l63 and other mutants (*wt* mutants) with 3lzm. The coordinate RMSDs are calculated from structures pre-aligned using all heavy atom coordinates. As described in the methods section unobserved, zero occupancy. High temperature factor and "symmetry" atoms are not used for comparison. If alternate locations of atoms are given, then the atoms most similar to the comparative structure are used. The lines are the regression lines and their 95% confidence bands. For a discussion of the outliers see the text.**

If the mutations have a large local effect on the structure, then one would expect the largest coordinate differences at the shortest distance to the site of mutation and vice versa the smallest coordinate differences furthest away from the site of mutation. However, when I plot the coordinate difference between mutants and wild types as a function of distance from the site of mutation (Figure 28), then I do not observe this pattern. While the largest coordinate difference is in the vicinity of the site of mutation, the trend is towards larger coordinate differences, as one move away from the site of mutation. One could try to explain the largest coordinate differences being furthest away because of a hinge motion in the protein, but I observe a reverse trend for the *wt* structures (green line in Figure 28), which does not support the argument. The best explanation for the data is that, while mutations do cause nearby conformational changes, the most pronounced effect of a single point mutation is that it causes a wider and redistributed conformational sampling. Because the intrinsic motion of T4L is a hinge motion, the coordinates furthest away from the center of the hinge at the center of the protein will be most displaced at the end points of the hinge motion.

In the next section I will show that very large RMSDs are observed, when indeed T4L is locked in one of the two end point conformations of the hinge motion.

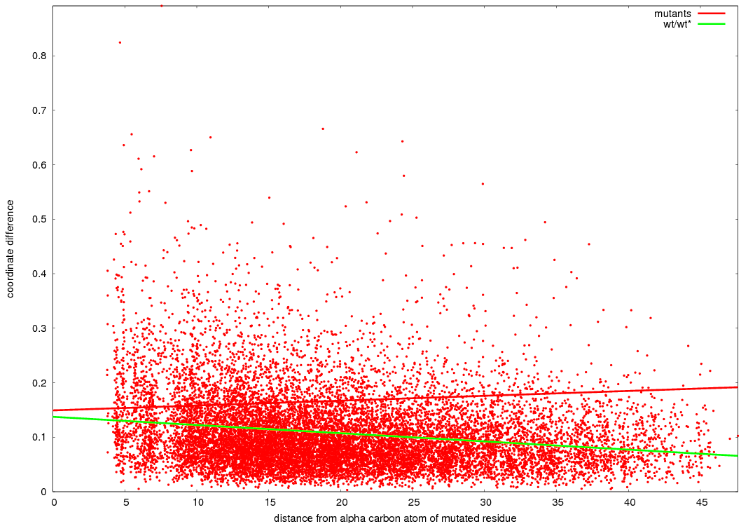


Figure 28 – A scatter plot of the distance between coordinates of a mutant structure aligned with a *wt* structure as a function of the distance from the mutated residue. The sequence similar structures (wt and wt*) were compared against each other. A sampling similar to those of the mutants was carried out for the wt/wt* structures. The data points of the wt/wt* comparisons are not shown. The regression lines for the mutant structures and wt/wt* are shown in red and green, respectively. When performing the analysis on all single point mutants, the mutations do not have a larger effect when they are in the vicinity of the mutation as compared to when they are at a distance from the mutation. The fact that the coordinates differ more when further from the site of mutation in the mutants might be due to a hinge motion, which is more pronounced in the mutants.

7.3.5.1 T4L special cases

Grey points of Figure 26 are special cases. Some of these are high RMSDs due to the conformation of T4L being locked at the end point (closed conformation) of the hinge motion. The hinge gap between residues 42 and 121 can be bridged by mutation to Cysteine. It has also been bridged by mutation to Histidine and by adding Cobalt, Nickel and Zinc respectively. In the absence of a bridging ion (257l) the T21H/T142H-mutant is not very different from the average wild type T4L structure. That the mere presence of an ion can lock T4L into an extreme conformation shows that the energy required to achieve this conformation is low, and it is therefore likely that the conformation is sampled in solution, although the population of this state is low in the absence of an ion.

Another special case is the A129W mutation. Ala129 is a buried residue in the C-terminal core of T4L. Mutating Ala129 to Trp causes steric clashes that cannot be avoided only by side chain reconfiguration, but also requires backbone movement. Another mutation which causes significant steric clashes is V87M. On its own, it has no major effect (1cu3), but in combination with L84M, L91M, L99M, V111M, L118M, L121M, and L133M, the main chain is forced to move to make room for the large side chain of Methionine. Vice versa the seven mutations of Leu and Val111 to Met has no significant effect (1l0k), unless Val87 is also mutated to Met (1lwg,1lwk,1lpy). Other mutants that are treated separately include the double Gly-to-Ala mutant 1ssy (G28A,I29A,G30A), the poly-Ala mutants 1l75 (127-133) and 192l (40-48,127-132).

A few structures have some awkward Gly107/Gly110/Gly113 ϕ/ψ angles as can be seen from the Ramachandran plots for these three residues (Figure 30). This is not surprising, as the F helix (residues 109-113) – in which the Glycine residues are located – has been observed to be mobile as evidenced by high thermal factors from crystallographic experiments(Morton and Matthews 1995), NMR peak broadening and methyl Carr-Purcell-Meiboom-Gill (CPMG) relaxation dispersion NMR experiments(Mulder, Mittermaier et al. 2001; Mulder, Hon et al. 2002), a wide range of different chemical shifts(Bouvignies, Vallurupalli et al. 2011), NMA (next chapter) and MD simulations(Arnold and Ornstein 1992). The structures with unusual Gly107/Gly110/Gly113 ϕ/ψ angles are mutant L99G (1qud), polar cavity mutants L99A/M102Q (3htf , 3htg, 3huk), mutant L99A/M102L (3dke), M102A/M106A (252l) and mutant L99A (3hh6). Here I seek to explain those Ramachandran outliers. L99A creates a buried cavity, whereas L99G (1qud) creates a solvent accessible declivity(Wray, Baase et al. 1999), which explains the alternate conformation of the F helix (residues 109-113) in 1qud (L99G). Two alternate conformations of helix F are present in 3hh6 (L99A). Of the two, the atoms of the one with the unusual conformation of residues 108-114 only have occupancies of 23%. Similarly in 252l (M102A/M106A), residues 106-115 come in two flavors. The electron density is not provided for 3dke (L99A/M102L), but perhaps it too has two conformations of the F helix of which only one has been modeled. The M102L mutation in 3dke (L99A/M102L) was only carried out to avoid possible artifacts in the electron density due to a selenomethionine (SeMet102) being close to the L99A cavity.(Liu, Quillin et al. 2008) It may be that the M102L mutation in combination with the L99A mutation unexpectedly causes the deformation of the F helix. Finally, the ligand present in the polar cavity of a L99A/M102Q mutant affects the conformation of the 107-113 turn. An apolar cavity was engineered into T4L through mutation of Leu99 to Ala. Despite the mutation creating a ≈150Å^3 cavity, the L99A mutant and the T4L *wt* structure are almost identical.(Eriksson, Baase et al. 1992) When the cavity is occupied by a nonpolar ligand, the structure is also very similar to the *wt* structure. However, if the apolar cavity is occupied by a polarized ligand, then the conformation in the C-terminal domain changes significantly. Those cases are colored with dark gray in Figure 26. The polar ligands that perturb the T4L L99A structure are shown in Figure 29. This is serves as an example, that predicting conformational changes is difficult, because the conformation can be heavily dependent on the nature of the ligand rather than the position and type of mutation. In the next chapter I analyze all structures in the protein data bank (PDB), and among other things I look into the structural effect different ligands have on sequence identical structures. The number of HEWL and T4L structures is not sufficient for this automated large scale analysis, and the HEWL and T4L structures present a very limited number of structures not solved by molecular replacement.

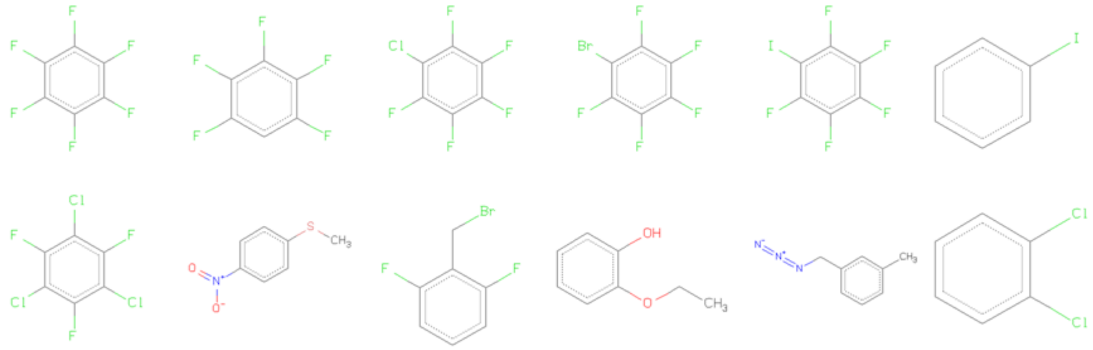


Figure 29 – Chemical structures of polar ligands that change the main chain conformation, when they are present in the apolar cavity of the wt*+L99A mutants. Non-polar ligands do not change the backbone conformation of the T4L wt*+L99A mutant.

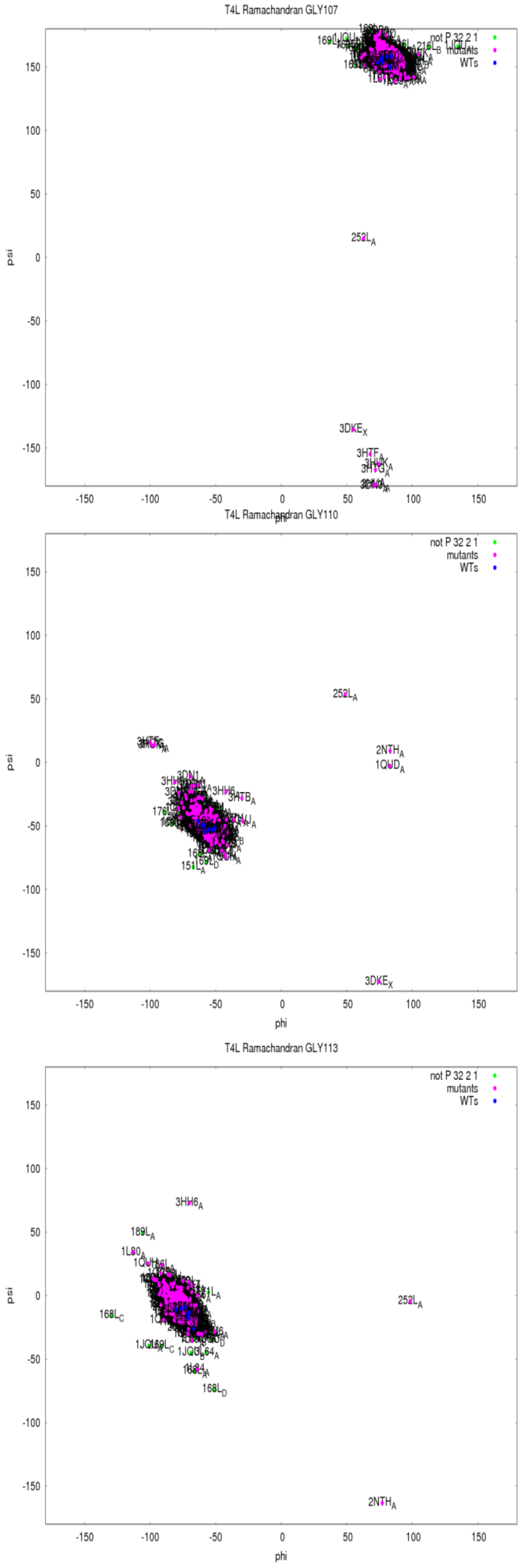


Figure 30 – Ramachandran outliers in the L99A/M102Q mutants of T4L. Gly107 (left), Gly110 (middle), Gly113 (right). The outliers are 252l:A, 2nth:A, 1qud:A.

7.3.6 Variation of structure dependent results

That the variation between structures measured in terms of C_α RMSD and differences in side chain dihedrals is mostly dependent on the origin of the structure (i.e. author/lab and starting model if molecular replacement) is unsettling in itself. More disturbing is the fact that structure based calculations based on seemingly identical structures yield very different results when calculating stability changes ($\Delta\Delta G$ values) of mutants (Figure 31) and pK_a values (Figure 32). This serves as a warning that computational results should be interpreted with care. While being qualitatively correct, the computed results can be quantitatively incorrect. One should always be aware of the error associated with computational results.

As highlighted by figure 32, the experimental pK_a value on the x-axis is larger than the minimum and smaller than the maximum of the ensemble of calculated values; the experimental value is sampled so to speak. It would therefore make sense to always perform structure based calculations on an ensemble of structures rather than on individual structures. The ensemble could be generated by NMA in those cases, when a sufficient number of X-ray structures are not available. It would be interesting to create an ensemble of structures with NMA and do structure based calculations again to see if experimental values are sampled using the NMA generated ensemble as an input. Alternatively the ensemble could be generated from experimental data by using one of two new methods for generation of hidden/invisible conformations through either the combination of chemical shifts from NMR CPMG relaxation dispersion experiments and a version of the structure prediction program Rosetta, which can utilize chemical shifts (CS-Rosetta)(Bouvignies, Vallurupalli et al. 2011) or directly from x-ray data.(Fraser, Clarkson et al. 2009)

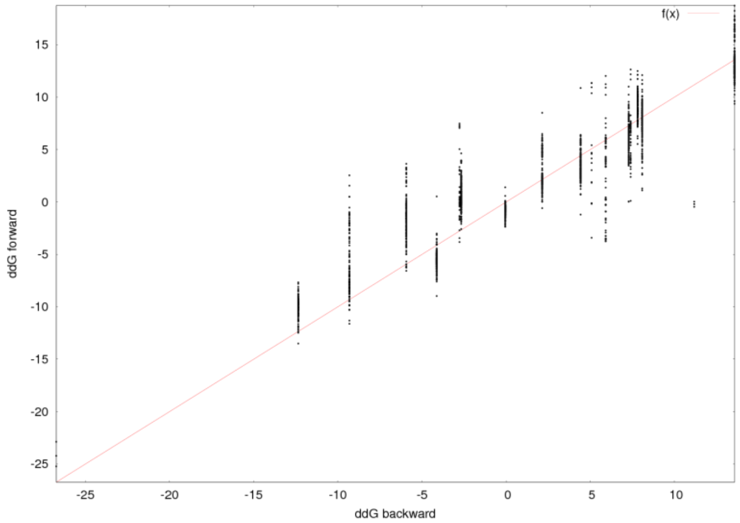


Figure 31 - Plot showing the correlation between UFFBAPS(Johnston, Søndergaard et al. 2011) calculated ΔΔG values; ΔΔG for a mutation (ddG forward, kJ/mol) and ΔΔG for the reverse mutation (ddG backward, kJ/mol). Miscellaneous HEWL *wt* structures are used for calculating ΔΔG forward, whereas only 1 mutant structure is available for the reverse calculation. The large variation in ΔΔG forward shows stability change calculations to be very dependent on the input structure.

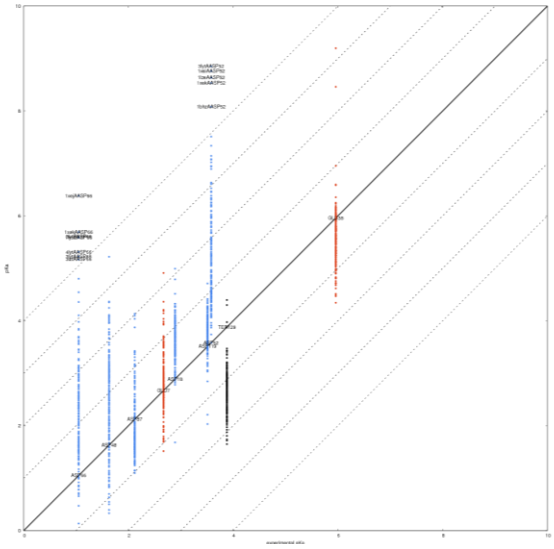


Figure 32 – Variation of pK_a values calculated from different HEWL structures. On the x-axis is the experimental pK_a value. On the y-axis is the calculated pK_a value. Asp and Glu residues are shown in blue and red, respectively. The C-terminal residue 129 is shown in black. A large variation in calculated pK_a values is observed for Asp66 (lowest experimental Asp pK_a) and Asp52 (highest experimental Asp pK_a). Among the severe outliers (those more than 4 pK_a units from the experimental pK_a value and more than 3 pK_a units from the average of the calculated pK_a values) are the usual suspects. 1) The dehydrated structures 1xei, 1xej, 1xek. 2) The structures 3lyt, 4lyt, 6lyt exposed to radiation induced decay. 3) 1lze, which is interacting with tetra-NAG, respectively. 4) 1bhz, which is solved at a 3.9Å resolution. 5) 2a6u, which is solved by means of powder diffraction. 6) 1e8l, which is solved by NMR spectroscopy. 7) The B chain of the P 21 structure 1lys.

7.4 Conclusions

This chapter has shown a linear relationship between the RMSD of two structures and their difference in Matthews coefficient. This result establishes that, upon structural comparison both the space group and also the level of hydration need to be taken into consideration.

Additionally, the chapter showed that the mutant structures of HEWL are more similar to the starting model used for molecular replacement than any other *wt* structure. This highlights the importance of strictly comparing mutant structures with their starting model upon investigating the effect of single point mutations. It shows that the exact coordinates of a structure deposited in the PDB is the results of lab methods such as starting model, buffering agents and molecular modeling software rather than just being the result of experimental observations.

At the 7th protein structure prediction competition "Critical Assessment of Techniques for Protein Structure Prediction" (CASP7), a refinement category of high resolution models was introduced. If information about crystal contacts are not made available to predictors, then it is difficult with force field based methods to model side chain conformations in the vicinity of crystal contacts more accurately than what can be obtained by simply copying the best template(MacCallum, Hua et al. 2009). As I have shown here, differences in crystal contacts – caused by different space groups or different levels of hydration – is the key to any significant structural heterogeneity. Mutations, ligands and different starting models, are only of secondary importance.

Figure 27 shows that the χ_1 angles of surface residues are more variable than those of buried residues. Therefore surface χ_1 angles should be excluded or down weighted when analyzing χ_1 angles of CASP models, which is currently not the case.(Randy and Gayatri 2007)

It is impossible to tell if the similarity between structures – whether solved in the same lab and/or from the same starting model – is due to similar experimental methods or to a phase bias. However, this chapter leaves no doubt that these similarities in experimental procedure contribute more to structure similarity than physiochemical differences contribute to structural differences.

8 Chapter 3 - Protein structure variation in the entire PDB

8.1 Introduction

The experimental X-ray structures deposited in the PDB are human interpretations of the electron density observed in an X-ray diffraction experiment.(Rupp 2009) Electron density of distinct conformations will be visible at higher resolution(Jelsch, Teeter et al. 2000; Schmidt, Teeter et al. 2011), but most often the electron density is itself an average of the ensemble of conformations that the protein occupies under the conditions in the crystal.(Fox, Evans et al. 1986) Conformational differences between PDB protein structures can thus be due to 1) the structural heterogeneity inherent in any ensemble(Volkman, Lipson et al. 2001; Fraser, Clarkson et al. 2009; Bouvignies, Vallurupalli et al. 2011), 2) differences in how the electron density was interpreted(Jones, Zou et al. 1991; Kleywegt and Jones 1996; Jones and Kleywegt 2007), or 3) genuine differences in structures/ensembles that are due to physiochemical differences such as pH(Zand, Agrawal et al. 1971; Tews, Findeisen et al. 2005; Roche, Bressanelli et al. 2006; Kukic, Farrell et al. 2009) and temperature(Frauenfelder, Petsko et al. 1979; Fisher, Colen et al. 1981) and/or mutations(Eriksson, Baase et al. 1992). In the previous chapter I studied the structural heterogeneity in an ensemble of HEWL and T4L structures and identified the importance of the starting model used for molecular replacement to the final deposited structure. In this chapter I primarily investigate 1) whether the overall conformational differences observed between structures deposited in the PDB can be explained by mutations and 2) whether the structural effect of a mutation is local or propagates from the site of mutation. I find that a single point mutation in general has no effect on the overall conformation of a protein and that a single point mutation only changes the conformation of the protein in the nearby vicinity (< 10Å) of the mutation.

8.1.1 Protein structure comparison

Protein structure comparison is relevant when studying the structural effect of the binding of a ligand(Kelly, Sielecki et al. 1979; Chothia, Lesk et al. 1986; Milburn, Prive et al. 1991; Mozzarelli, Rivetti et al. 1991; Rini, Schulze-Gahmen et al. 1992; Ostermann, Waschipky et al. 2000), when performing sequence independent protein structure alignment and clustering(Shindyalov and Bourne 1998; Ortiz, Strauss et al. 2002; Ye and Godzik 2004; Zhang and Skolnick 2005), when analyzing protein equilibrium fluctuations(McCammon, Gelin et al. 1977) and protein folding(Lindorff-Larsen, Piana et al. 2011) with molecular dynamics simulations, when benchmarking and analyzing computational methods for structure prediction(Schulz, Barry et al. 1974; Delbaere, Brayer et al. 1979; Chothia, Lesk et al. 1986; Schueler-Furman, Wang et al. 2005), when analyzing the extent of conformational sampling carried out by a protein at equilibrium(Lindorff-Larsen and Ferkinghoff-Borg 2009), and most importantly when analyzing the structural effect of mutations.

Mutations can perturb protein structures in multiple ways. Mutations can change the conformation by disruption or introduction of hydrogen bond networks(Alber, Dao-pin et al. 1987; Sprang, Standing et al. 1987), salt bridges(Gibbs, Moody et al. 1990), disulfide bridges(Pjura, Matsumura et al. 1990; Miller, Mande et al. 1995) helix dipole interactions(Nicholson, Becktel et al. 1988) and hydrophobic cavities(Eriksson, Baase et al. 1992; Korkegian, Black et al. 2005). Mutations can also indirectly alter the conformation observed by X-ray diffraction by altering the rate of conformational change between two conformational states(Bhabha, Lee et al. 2011) and thus the size of the population of the two states and ultimately the observed electron density and the reported conformation.(Bouvignies, Vallurupalli et al. 2011)

When performing *de novo* protein structure prediction, a C_α RMSD smaller than 1.5Å – which can be achieved for small protein domains (<85 residues) – is considered high resolution.(Bradley, Misura et al. 2005) In this chapter I show that large single domain proteins from different space groups with at least 95% sequence similarity rarely display C_α RMSDs larger than 1.5Å. In fact the average C_α RMSD of an ensemble of sequence identical structures rarely exceeds 0.5Å. Therefore there is still a gap between the accuracy of the experimental determination and that of the computational prediction of a structure. However, this gap will narrow, as computational methods of conformational sampling and free energy calculations improve.(Bradley, Misura et al. 2005)

The study presented in this chapter is somewhat similar to that of Kosloff and Kolodny(Kosloff and Kolodny 2008), which identified sequence-similar but structure-dissimilar protein pairs in the PDB and explained those structural differences on an individual basis. I am however not interested in RMSD outliers, but rather I want to know, which properties in general cause structural differences.

8.1.2 RMSD as a measure of structural similarity

It has previously been shown that the RMSD between two structures is dependent on their size; i.e. their radius of gyration. This result however was established for sequence different proteins.(Cohen and Sternberg 1980; Reva, Finkelstein et al. 1998) Because I compare proteins of different size, it is very important for me to use a measure, which is not size dependent. One option is to use a size-independent RMSD.(Maiorov and Crippen 1995; Betancourt and Skolnick 2001; Carugo and Pongor 2001) Another option is to show that the RMSD is size-independent for my dataset. In section 8.3.4.5 I show that the RMSD for sequence identical structures is not dependent on their radius of gyration. It is therefore a valid approach to use the RMSD as a measure of structural similarity for proteins of different size.

8.1.3 Structure of the chapter

In this chapter I seek to quantify the contribution of ensemble heterogeneity to conformational differences, and identify the properties that contribute to conformational differences by performing a pair wise comparison of a large set of protein structures of different sizes, folds and functions.

I set out to determine which factors cause variation between otherwise sequence identical proteins. For example it is important to be able to compare an apo structure and a ligand bound structure and know whether the observed differences in the atomic coordinates of the protein is due to the binding of the ligand or simply a result of ensemble heterogeneity or differences in how the electron density was interpreted. Specifically I examine to what extent the variation between structures can be explained by mutations (section 8.3.3.2), the presence/absence of a ligand, the presence of differing ligands and frequently reported physiochemical properties such as pH (8.3.4.4) and temperature (8.3.4.1). The results show that there are some physiochemical properties that have a detectable influence on the conformation of a protein. To summarize, the two main questions I seek to answer in this chapter are:

1) Why are sequence similar structure pairs sometimes structurally different? Can this be attributed to something else than a point mutation, which one should take into consideration, when attributing structural differences to point mutations? This analysis builds on the results from the analysis of HEWL and T4L (chapter 7) and examines whether the conclusions reached from this study is likely to be true for a larger subset of the protein structures in the PDB. Whereas the analysis in the previous chapter focused on the relationship of structures through molecular replacement, this chapter will focus on the analysis of whether structural differences are observed in the case of mutations and physiochemical differences. This is impossible to answer for HEWL, since less than 49 of the 349 HEWL structures in the PDB are mutant structures. For T4L it is complicated by the fact that 467 of 530 T4L structures in the PDB are solved by the research group of Brian Matthews.(Baase, Liu et al. 2010) The given conformation of a solved T4L structure would depend too much on whether it had been solved in the lab of Brian Matthews and/or by molecular replacement using a starting model from Brian Matthews.

2) Does the structural effect of a mutation propagate through a protein originating from the site of mutation? Can the structural effect of a mutation be quantified or do other properties

contribute to the structural difference between a structure pair and inhibit any quantification due to "background noise"? I present my answers to these questions in section 8.3.5.

8.2 Methods

A flow diagram for finding comparable structures is shown in Figure 33. Before the comparison of structures is carried out, the pool of structures from the PDB is reduced. This is explained in section 8.2.1.

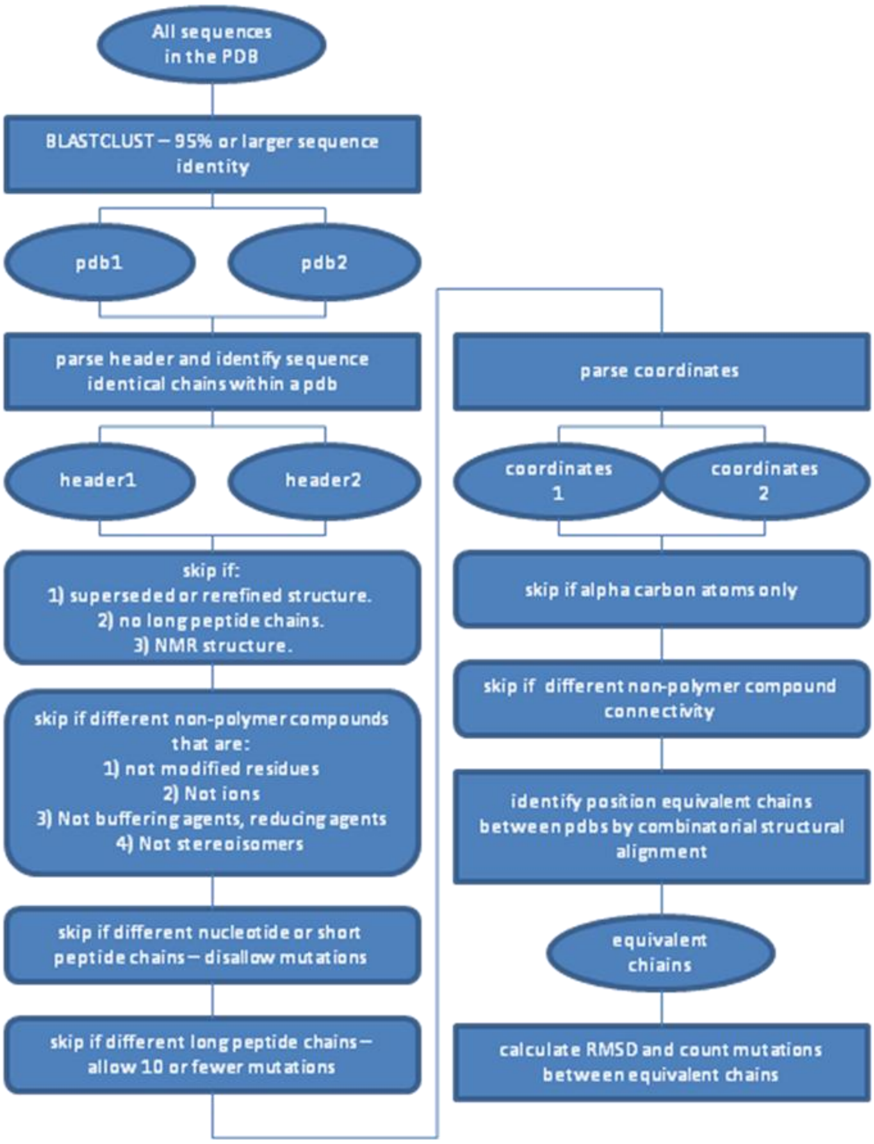

All sequences in the PDB
BLASTCLUST – 95% or larger sequence identity
pdb1
pdb2
parse header and identify sequence identical chains within a pdb
header1
header2
skip if:
1) superseded or rerefined structure.
2) no long peptide chains.
3) NMR structure.
skip if different non-polymer compounds that are:
1) not modified residues
2) Not ions
3) Not buffering agents, reducing agents
4) Not stereoisomers
skip if different nucleotide or short peptide chains – disallow mutations
skip if different long peptide chains – allow 10 or fewer mutations
parse coordinates
coordinates 1
coordinates 2
skip if alpha carbon atoms only
skip if different non-polymer compound connectivity
identify position equivalent chains between pdbs by combinatorial structural alignment
equivalent chains
calculate RMSD and count mutations between equivalent chains

Figure 33 – Flow diagram of the protein structure comparison algorithm.

8.2.1 Exclusion of PDB files

183 PDB files were excluded due to errors in the files.

Other PDB files that are excluded are those for which only the C_α atoms are modeled (e.g. 1tt9), those with REMARK 0 records in them (i.e. re-refined structures; e.g. 2pfd, which is a rerefined model of 1tt9), those that are superseded or obsolete, those that were not generated from X-ray diffraction data and those with multiple models in them.

An additional 100 to 150 structure pairs are excluded, because the application of any type of combination of the given symmetry operators yields widely different biological units as measured by RMSD. These are not errors per se, but I prefer to exclude a very limited number of true positive RMSD outliers instead of including false positives.

As already stated I exclude T4L and HEWL from the dataset to avoid redundancy and because they are analyzed separately in chapter 7. These two proteins are heavily represented in the PDB with 522 and 349 structures respectively.

8.2.2 Selection of protein structure pairs

8.2.2.1 Sequence

Only if two chains have more than 95% sequence identity are they compared. The clustering of chains with 95% sequence similarity was carried out with BLASTCLUST(Altschul, Gish et al. 1990) instead of CD-HIT(Li and Godzik 2006), despite CD-HIT being the default algorithm used by the PDB for clustering structures by sequence similarity. The choice was made to avoid peptide chains with terminal differences not being compared. By choosing to ignore terminal differences it is also ensured that proteins, for which any N-terminal signal peptides have been included in the sequence by mistake (e.g. HEWL structures 1lsy, 1lsz), are not excluded from comparison. Terminal differences are often due to cloning artifacts and expression tags that have little effect on the overall structure of the protein. Figure 34 shows sequences of peptide chains that are clustered differently with BLASTCLUST and CD-HIT. The RMSD between those structures is small, despite the differences in the N-terminal sequence.

An example of sequence differences at the terminus that cause large structural differences is observed in for example the T7 helicase (PDB IDs 1cr0, 1e0j) and the C-terminal domain of TonB (PDB IDs 1ihr, 1u07). Those are examples of domain swapping. Domain swapping is often caused by a single point mutation. A better way of handling the comparison of domain swapped proteins would be to compare the domains to each other individually. Frequently domain swapping involves a hinge motion(Liu and Eisenberg 2002), and I predict that normal mode analysis will be very successful in predicting the motion between two domain swapped conformations.

```
1s9q_A         MGSSHHHHHHSSGLVPRGSHMPAKKPYNKIVSHLLVAEPEKIYAMPDPTVPDSDIKALTT 60
1s9q_B         MGSSHHHHHHSSGLVPRGSHMPAKKPYNKIVSHLLVAEPEKIYAMPDPTVPDSDIKALTT 60
1tfc_A         MGSSHHHHHHSSGLVPRGSHMPAKKPYNKIVSHLLVAEPEKIYAMPDPTVPDSDIKALTT 60
1tfc_B         MGSSHHHHHHSSGLVPRGSHMPAKKPYNKIVSHLLVAEPEKIYAMPDPTVPDSDIKALTT 60
1vjb_A         MGSSHHHHHHSSGLVPRGSHMPAKKPYNKIVSHLLVAEPEKIYAMPDPTVPDSDIKALTT 60
1vjb_B         MGSSHHHHHHSSGLVPRGSHMPAKKPYNKIVSHLLVAEPEKIYAMPDPTVPDSDIKALTT 60
2p7a_A         MGSSHHHHHHSSGLVPRGSHMPAKKPYNKIVSHLLVAEPEKIYAMPDPTVPDSDIKALTT 60
2p7g_A         MGSSHHHHHHSSGLVPRGSHMPAKKPYNKIVSHLLVAEPEKIYAMPDPTVPDSDIKALTT 60
2p7z_A         MGSSHHHHHHSSGLVPRGSHMPAKKPYNKIVSHLLVAEPEKIYAMPDPTVPDSDIKALTT 60
2gpv_F         ---------------------PAKKPYNKIVSHLLVAEPEKIYAMPDPTVPDSDIKALTT 39
2gpv_E         ---------------------PAKKPYNKIVSHLLVAEPEKIYAMPDPTVPDSDIKALTT 39
2gpv_D         ---------------------PAKKPYNKIVSHLLVAEPEKIYAMPDPTVPDSDIKALTT 39
2gpv_C         ---------------------PAKKPYNKIVSHLLVAEPEKIYAMPDPTVPDSDIKALTT 39
2gpv_B         ---------------------PAKKPYNKIVSHLLVAEPEKIYAMPDPTVPDSDIKALTT 39
2gpv_A         ---------------------PAKKPYNKIVSHLLVAEPEKIYAMPDPTVPDSDIKALTT 39
2gpu_A         ---------------------PAKKPYNKIVSHLLVAEPEKIYAMPDPTVPDSDIKALTT 39
2gpp_B         ---------------------PAKKPYNKIVSHLLVAEPEKIYAMPDPTVPDSDIKALTT 39
2gpp_A         ---------------------PAKKPYNKIVSHLLVAEPEKIYAMPDPTVPDSDIKALTT 39
2gpo_A         ---------------------PAKKPYNKIVSHLLVAEPEKIYAMPDPTVPDSDIKALTT 39
2gp7_D         ---------------------PAKKPYNKIVSHLLVAEPEKIYAMPDPTVPDSDIKALTT 39
2gp7_C         ---------------------PAKKPYNKIVSHLLVAEPEKIYAMPDPTVPDSDIKALTT 39
2gp7_B         ---------------------PAKKPYNKIVSHLLVAEPEKIYAMPDPTVPDSDIKALTT 39
2gp7_A         ---------------------PAKKPYNKIVSHLLVAEPEKIYAMPDPTVPDSDIKALTT 39
1kv6_B         ---------------------PAKKPYNKIVSHLLVAEPEKIYAMPDPTVPDSDIKALTT 39
1kv6_A         ---------------------PAKKPYNKIVSHLLVAEPEKIYAMPDPTVPDSDIKALTT 39
1s9p_D         ------------------------KPYNKIVSHLLVAEPEKIYAMPDPTVPDSDIKALTT 36
1s9p_C         ------------------------KPYNKIVSHLLVAEPEKIYAMPDPTVPDSDIKALTT 36
1s9p_B         ------------------------KPYNKIVSHLLVAEPEKIYAMPDPTVPDSDIKALTT 36
1s9p_A         ------------------------KPYNKIVSHLLVAEPEKIYAMPDPTVPDSDIKALTT 36
2ewp_E         -------------------------PYNKIVSHLLVAEPEKIYAMPDPTVPDSDIKALTT 35
2ewp_D         -------------------------PYNKIVSHLLVAEPEKIYAMPDPTVPDSDIKALTT 35
2ewp_C         -------------------------PYNKIVSHLLVAEPEKIYAMPDPTVPDSDIKALTT 35
2ewp_B         -------------------------PYNKIVSHLLVAEPEKIYAMPDPTVPDSDIKALTT 35
2e2r_A         --------LGSPEFLNPQLVQPAKKPYNKIVSHLLVAEPEKIYAMPDPTVPDSDIKALTT 52
2zas_A         --------LGSPEFLNPQLVQPAKKPYNKIVSHLLVAEPEKIYAMPDPTVPDSDIKALTT 52
2zbs_A         --------LGSPEFLNPQLVQPAKKPYNKIVSHLLVAEPEKIYAMPDPTVPDSDIKALTT 52
2zkc_A         --------LGSPEFLNPQLVQPAKKPYNKIVSHLLVAEPEKIYAMPDPTVPDSDIKALTT 52
2ewp_A         -------------------------PYNKIVSHLLVAEPEKIYAMPDPTVPDSDIKALTT 35
```

Figure 34 – N-terminal sequences of Estrogen Related Receptor Gamma ligand binding domain. When BLASTCLUST 95% sequence identity is used, then all of the 38 sequences are clustered together. When CD-HIT 95% sequence identity is used, then the top 9 and bottom 29 sequences are clustered separately. MGSSHHHHHHSSGLVPRGSHM is a cloning artifact and an expression tag, which is why it should be ignored and not be the cause of sequences being clustered separately.

The number of mutations – or rather differing residues among two chains – is allowed to range between 0 and 10 per chain. Chains with more than 10 mutations between them are not paired. Insertions and deletions at the N- and C-terminus are ignored when doing sequential alignment; i.e. terminal residues present in only one of the two chains being compared are ignored. However, chains with a difference in length exceeding 25 residues are not paired. Polypeptide chains with less than 50 residues are treated as peptide ligands; i.e. differences in length and sequence are strictly not allowed and the chains are not included in the calculation of the RMSD between the two biomolecules being compared. The cutoff of residues is quite high, as some proteins are shorter than 50 residues, but I prefer to exclude a few real protein structures over the inclusion of many peptide ligands.

8.2.2.2 Structure

PDB files are paired if all of the long peptide chains in one PDB file have a sequence similar chain in the other PDB file, and the nucleotide chains, short peptide chains and non-polymer compounds are identical. I go into more detail about the process of ligand comparison in section 8.2.2.4. Once two PDB files have been paired, then structurally equivalent chains are identified and an RMSD is calculated for those structurally equivalent chains that make up a biomolecule. Chains are considered structurally equivalent if the RMSD falls below empirically determined thresholds of 4.69Å for monomers, 6.19Å for dimers, 6.21Å for trimers and 6.25Å for biomolecules with four or more chains.

If the biological units provided by the REMARK350 transformations are different for a pair of structures, I use biological units as determined by PISA(Krissinel and Henrick 2007).

If neither the REMARK350 records nor PISA provides identical biological units, then all combinations of chains and space group symmetry operators in the REMARK290 records (and 27 translations) are tried, until the RMSD falls below the threshold. The quaternary structures obtained by this method might not be biological units, but in terms of RMSD that is of no importance as shown in figure 35, if the space group and unit cell dimensions of the two PDB files are identical. When the space groups and unit cell dimensions are identical, then RMSD is independent of the absolute positioning of individual chains. If the unit cell dimensions of two structures are different, then that will contribute to a higher RMSD. For instance the unit cell dimensions of the non-falsified but retracted PDB entry 179l (T4L) were wrong.(Sheffler and Baker 2009; Tronrud and Matthews 2009) This transformation problem could in theory apply to all biomolecules achieved by unit cell transformations. I have however compared the RMSD of transformed biological units and those with non crystallographic symmetry, and I find no difference between the two sets of pairs of biomolecules.

If no identical quarternary structures can be found, then it's concluded that the biological units are different. These PDB pairs are excluded from the dataset. Some cases include different spacing due to different unit cell side lengths (PDB IDs 1hno, 1phj and 2reb,1u94) and different dimer interfaces that are not related through symmetry operations (PDB IDs 1lcu, 2q31).

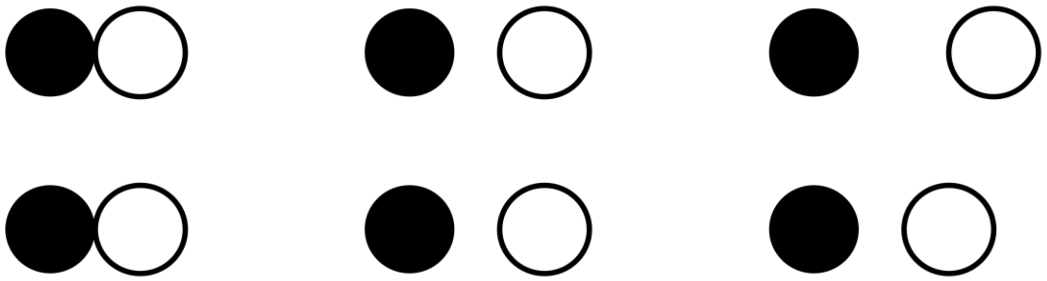


Figure 35 – The RMSD between the top and bottom biological units (left column) is the same as that between the top and bottom quarternary structures (middle column), in which a redundant transformation is applied to a subset of the chains (white circle) constituting the biological unit, if the space group and unit cell dimensions in the two PDB files are identical. If the unit cell dimensions are different, then applying unit cell translations will change the RMSD.

8.2.2.3 Residues

Selenomethionines are not treated as modified residues but as their standard parent residue Methionine. UNK, ASX and GLX residues are treated as unknown residues when comparing sequences. In the case of microheterogeneity in peptide sequences due to alternate conformations the residue with its name present in the SEQRES and MODRES records is the one that is used for structural comparison (e.g. LLP instead of PLZ in 2okj).

A residue is used for RMSD calculation, if that residue is present in both biomolecules being compared. Otherwise the structural alignment is carried out for the remainder of the observed/modeled residues. Zero occupancy residues are included in the structural comparison.

The super positioning of structures is based on the quaternion method (code by David J. Heisterberg from The Ohio Supercomputer Center, 1990, unpublished results). C_α atoms are used for the super positioning of two structures and the concomitant calculation of RMSD between the two structures. Side chain atoms are not used, because mutated residues are compared occasionally. For speed purposes it's not all of the heavy backbone atoms that are used for calculation of the RMSD between biomolecules. Instead just C_α atoms are used.

8.2.2.4 Ligands

Different ligands could cause different conformations. Therefore it is a requirement that peptide and nucleotide ligands have identical sequences. Furthermore the identity and connectivity of other ligands (e.g. saccharides) have to be identical. And the site of glycosylation has to be identical between the two biomolecules being compared.

Proteins are only compared, if their non-ion ligands are identical. The check is initially carried out by comparing chemical component IDs. The number of ligands is not considered. Only their identity is taken into consideration. Modified residues are not treated as ligands. A list of selected ions and other common solutes are ignored upon comparison of hetero compound IDs (see appendix).

If a ligand has two or more alternative identities, then both ligands are used for comparison of ligands between biomolecules. Only one of multiple alternative ligand identities of one structure has to match that of the other structure it is being compared to.

If a ligand has two or more alternative connectivities, then both connectivities are used, when comparing the bonding pattern of ligands between two biomolecules.

8.2.3 RMSD as a measure of difference between a pair of pdb structures

The RMSD is a measure of the difference between a pair of PDB structures as explained in the previous chapter. The RMSDs that I calculated for the determining the structural effect of various factors were initially based on all heavy atoms, but I noticed that surface residues often have quite different conformations, which can be attributed to their flexibility or different space groups causing different crystal contacts. To eliminate this cause of an imprecise RMSD I decided to calculate the RMSD of heavy backbone atoms only. To further speed up the millions of structural alignments I only use C_α atoms when doing structural alignment prior to calculating the RMSD.

A small terminal random coil or a small domain involved in a hinge motion can increase the RMSD of two otherwise identical proteins; i.e. a local dissimilarity involving only a few residues can cause a large overall dissimilarity as measured by RMSD. The comparison of 1zal, 1ald and 1onn, 1k5h is two of many examples of this. The small structural differences might be caused by hetero compounds that are not modeled or different crystal contacts. 1zal contains tagatose-1,6-bisphosphate, which is modeled as two phosphate ions. Had P6T been present instead of PO4, then 1zal and 1ald would not have been paired. The data quality is of outmost importance in this project, and it varies between structures in the PDB.

To get a more realistic measurement of the difference between two structures, one could exclude or weigh lower any residues, which increase the RMSD significantly. The identification of those super flexible residues could be based on 1) an initial super positioning, 2) normal mode analysis, 3) X-ray temperature factors, 4) surface exposure – assuming surface residues have higher degrees of freedom than buried ones or 5) secondary structure element types; i.e. exclusion of random coils. If flexible residues connect two domains and constitute a hinge between the two domains, then the three latter methods are not viable for weighing or excluding residues upon calculation of the RMSD.

ϕ/ψ differences unlike coordinate differences do not propagate throughout the entire structure of a protein. For the same reason ϕ/ψ RMSDs report poorly on large concerted conformational changes. I therefore only use coordinate RMSD when analyzing the structural effect of a set of parameters (section 8.3.2-8.3.4).

8.2.4 Statistical methods

To determine if a given parameter has an influence on a protein structure, I calculate the mean RMSDs for different sub groups of the dataset. I use statistical methods to determine, if the mean RMSDs are identical.

One statistical method I make use of is Student's *t*-test. The assumptions of the *t*-test are(Zar 1998):

1) The variance between groups is homogenous (homoscedasticity).

2) The data in each group is normal distributed (i.e. symmetric and unimodal).

3) The population size of each group is identical (balanced).

When presenting the analysis of each parameter I will explain, whether the assumptions hold true or not and thus whether the *t*-test is reliable or not.

I use the *t*-test to determine if differences between RMSD means are significant or not. I also use it to test if correlations between RMSDs and various factors are significant ($r \neq 0$) or not ($r = 0$). Because I do not want to make assumptions about a positive or a negative correlation I use a two-tailed *t*-test instead of a one-tailed *t*-test. The *t*-value that I calculate for correlation coefficients, *r*, is heavily dependent on the number of observations, *n*.

$$t_{r=0} = r\sqrt{\frac{n-2}{1-r^2}} \tag{8-1}$$

A large number of data points will lead to a significant correlation despite a low covariance. Therefore I normalize the data when necessary. That involves finding the correlation between average RMSDs on the y-axis at discrete frequently occurring x-values ($n_{\text{RMSDs}} > 100$). Two assumptions when performing linear regression has to hold. The two assumptions are met in all cases, unless I state otherwise.

1) For each value of x the variances of y are similar.

2) The y values are normal distributed.

8.3 Results

I present a qualitative and quantitative analysis of the effect of various parameters on structure differences between protein structures. For categorical parameters such as space group difference and for discrete value parameters such as the number of mutations and the number of chains I present a qualitative analysis (sections 8.3.2 and 8.3.3). I determine whether each parameter has an effect or not, but I do not quantify the effect. Instead I break the dataset into subsets for which I can rule out the effect of the categorical parameters and for these subsets I do a quantitative analysis of the effect of continuous value parameters such as pH, temperature, solvent content and radius of gyration (section 8.3.4). In section 8.3.5 I shift my focus to the analysis of the structural effect of mutations and investigate, whether a mutation has a stronger effect in its vicinity or distal from the site of mutation.

8.3.1 The dataset

The protein data bank contains 74888 PDB files as of August 3rd 2011. 65321 of those are X-ray structures, of which 61069 contain at least one peptide sequence. In my analysis I look at pairs of structures that are sequence similar. My dataset contains 17401 PDB files that are sequence similar (more than 95% sequence identical as described in methods section 8.2.2.1) to at least one other PDB file. The asymmetric unit of a PDB file can contain one or more biomolecules. Therefore the total number of biomolecules, which are sequence similar, is even higher (75916) than the number of sequence similar PDB files. The total number of pairs of sequence similar biomolecules is 346,059. To avoid redundancy multiple biomolecules in one PDB file are not compared against each other. Some PDB files are paired with only each other, whereas others are paired with multiple other PDB files. 3fi5, for example, is a crystal structure of T4 lysozyme with 4 biomolecules in the asymmetric unit and it is paired with 1,628 biomolecules. Lysozyme and other model proteins account for a large fraction of the paired biomolecules (Table 7). To avoid redundancy in the dataset and because I analyze T4L and HEWL separately in chapter 7, I do not include them in the analysis in this chapter.

Protein	Number of structures	Number of structure pairs (one biological unit in asymmetric unit)	Percentage of all structure pairs
T4L	522	169,332	48.9%
HEWL	352	39,006	11.3%
Human lysozyme	203	35,910	10.4%
Carbonic anhydrase II	379	11,990	3.5%
Diphthine synthase	82	11,130	3.2%
Hemoglobin	197	8,930	2.6%
Ribonuclease A	186	8,742	2.5%
Myoglobin	229	7,656	2.2%
Superoxide dismutase	65	5,402	1.6%
Staphylococcal nuclease	109	5,256	1.5%
BPTI	82	4,290	1.2%

Table 7 – The most frequently occurring structure pairs in the dataset.

Most of the 346,059 pairs of biomolecules are single chain proteins (Table 8). The average length of overlapping sequence in monomeric biomolecule pairs is 174 residues. The average C_α RMSD of monomeric biomolecule pairs is 0.56Å.

Oligomeric state	Number of structure pairs
Monomeric	279,187
Dimeric	37,879
Trimeic	4,642
Tetrameric	19,527
Pentameric	165
Hexameric	1,958

Table 8 – Frequency of selected oligomeric states in the dataset.

The number of mutations in the 279,187 monomeric protein pairs is shown in Figure 36. It should be noted that two mutations mostly correspond to the comparison of single point mutants to each other and not the comparison of wild types and double mutants. Likewise the cases of 4 mutations are in many cases due to the comparison of double mutants to each other. 42,157 biomolecule pairs have 1 mutation. 4,823 of those mutations are engineered, whereas the cause of the remainder of the single point mutations is not described in the PDB files. A fourth of the single point mutations (1156) are to Alanine. Table 9 further summarizes the 4,823 engineered mutations.

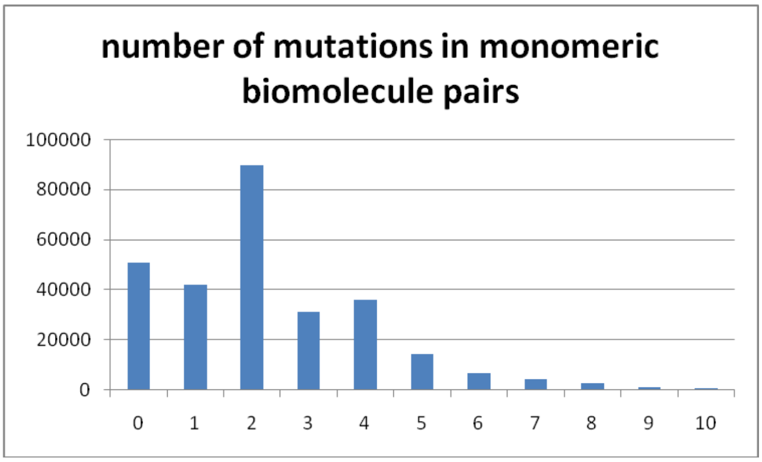


Figure 36 - Number of mutations in the 279,187 monomeric biomolecule pairs.

		mutated residue																				
		ALA	CYS	ASP	GLU	PHE	GLY	HIS	ILE	LYS	LEU	MET	ASN	PRO	GLN	ARG	SER	THR	VAL	TRP	TYR	total
wild type residue	ALA		2	5	2	13	3	7		2	10			10	5	2	28	3	6	2	3	103
	CYS	54		9			1	2		1							26					93
	ASP	55			32		1	7		10		1	209	11	4	1	4				2	337
	GLU	42		22			6	2		12			2		112	1	11		1			211
	PHE	74		1	1		37	19	9		46			3					51	66	30	337
	GLY	483		11	1								3	2	2	1	12	16	7			538
	HIS	126	39	37	1	94	140				13		29	1	24	1		13	96	12	18	644
	ILE	7	3			6					6	281	9	7	6				125	4	3	458
	LYS	16		3	19		7		1			1	1	7	7	19	3	1				85
	LEU	31			5	32	6	12	3	1		10	8	2		10	9		7	13		150
	MET	11				93			5	2	102						2	1	1			217
	ASN	14		106			20	55		2	52				2		2					253
	PRO	73	3	3	2	8	41	3	10		5					3	14	6	1		2	174
	GLN	15	1	5	25		3			8	13	1	5			1			1			78
	ARG	25	2		16	1	1	10		12		2			7		5			2		83
	SER	36	22	12		2	8				1		1			3		95	3			183
	THR	32	48	25	18	4	17	15	12		4	2	16			3	121		50	1		368
	VAL	15	6		14	11	9	4	17	3	17	6	9	5			4	11		2	4	137
	TRP	3				14	1	2			5					5					107	137
	TYR	44	1		4	74	37	19			43				3		1	1		9		236
	total	1156	127	239	140	352	338	157	57	53	318	304	292	48	172	50	242	147	349	111	169	4823

Table 9 - Substitution matrix of 4,823 engineered single point mutations in the dataset. The wild type residues are on the left column and the inserted residues are on the top row. Each cell in the table counts the number of mutations of that type. The bottom row and the right column show the totals. The table only contains sequence differences that are confirmed to be engineered mutants. For example *wt* structures from different organisms that differ by just one residue are not included.

8.3.2 Structural differences by category

At first I look at nominal value parameters with the purpose of splitting the dataset into categories – while keeping a sub dataset of a significant size – if these nominal value parameters have an effect on the average RMSD. The effect of each of the nominal value parameters are listed in table 10.

	$n_{\text{identical}}$	$n_{\text{different}}$	<RMSD>$_{\text{identical}}$ / Å	<RMSD>$_{\text{different}}$ / Å	*t*	*p*
space group	6426	2700	0.32	0.85	-24.7	< 0.01
ions	3467	2958	0.28	0.36	-9.60	< 0.01
authors	4617	1795	0.30	0.37	-6.89	< 0.01
pH	3350	1856	0.27	0.39	-10.8	< 0.01
T	3327	2060	0.29	0.35	-5.35	< 0.01
	n_{cryo}	n_{room}	<RMSD>$_{\text{cryo}}$	<RMSD>$_{\text{room}}$		
T	2736	1513	0.31	0.25	-6.12	< 0.01
	n_{present}	n_{absent}	<RMSD>$_{\text{present}}$	<RMSD>$_{\text{absent}}$		*p*
REMARK465	3582	2843	0.37	0.25	-13.7	< 0.01
REMARK470	2176	4099	0.35	0.31	-4.07	< 0.01

Table 10 – Statistics of the dataset when split into two categories according to nominal values. All tabulated nominal value parameters have a significant effect on the average RMSD, <RMSD>. The reported *t* and *p* values are the median values of 1000 statistical tests carried out on samples with identical population sizes, *n*. This was done to strengthen the reliability of the statistical *t*-test. In terms of significance level, *p*, all 1000 statistical tests showed the same result.

One of the main questions I am seeking to answer in this chapter is whether conformational differences between mutant structures in the PDB are due to the mutation or structural heterogeneity, experimental errors and other physiochemical parameters. For the further analysis I want to rule out the effect of categorical parameters, while retaining a dataset of a sufficient size. Space group differences seem to give rise to the largest difference in average RMSD. Therefore I solely look at structures with identical space groups from this point forward. The other parameters in the table all have a statistically significant effect, but it's smaller than that from space group differences. To maintain a dataset of a significant size I cannot split the dataset into further subcategories, and I have to assume that the other categorical parameters are normal distributed and independent of all other parameters in the further analysis in the next sections.

Interestingly whether the authors are different or not and whether residues (REMARK465) or atoms (REMARK470) are missing or not have an influence on the structure despite being properties that should not influence the structure (Table 10). The REMARK records merely express the quality of the structures observed. If they are present, it is not likely that other parts of the model rely on modeling rather than experimental observations. That an author difference causes a larger RMSD is a sign that some non-apparent factor has an effect on the crystallization procedure, the diffraction experiment or the interpretation of the experimentally observed electron density. The RMSD difference (0.30Å and 0.37Å) is significant as measured by Student's *t*-test ($p < 0.01$).

During this analysis of the PDB as a whole I was surprised by many C_α and heavy atom RMSDs between structures being close to zero despite mutations, ligand differences and space group differences. I noticed that structures from the same research group or solved by molecular replacement from the same starting model are often more similar to each other than structures from a different research group despite those dissimilarities between structures from different research groups not being explained by experimental conditions. I therefore decided to look at a subset of the PDB and investigate, whether structures from the same research group really are more similar to each other because of their common origin and not because of similar experimental conditions. For this I needed a large number of sequence similar structures solved at diverse experimental conditions by different research groups. I was preferably looking for monomeric proteins, so movements between chains did not have to be considered. Hen egg white lysozyme (HEWL) and bacteriophage T4 lysozyme (T4L) are two such proteins. I also analyzed the two proteins separately to discern structure variation due to physiochemical properties from variation due to intrinsic dynamics measured as an “average” over the ensemble of all structures. In chapter 7 I presented the results of my findings on these two structures.

I now look at categorical parameters, so I can determine, how the protein structures need to be sub categorized in order to have a data set of comparable protein structures that can be used to determine whether mutations have an effect or not.

8.3.2.1 Space groups and crystal contacts

Different crystal contacts cause different conformations as seen in table 10, which holds the statistics of the comparison of the RMSD between proteins from identical and non-identical space groups. The conformation of a protein is dependent ($p < 0.01$) on whether the space groups are identical ($<RMSD>_{identical}$ = 0.32Å) or different ($<RMSD>_{different}$ = .85Å). Thus it would make sense only to compare proteins from identical space groups. It has already been shown that the average resolution is independent of space group (Rupp 2009), so there is no need to split the dataset into subsets according to space group in case the resolution is a parameter to consider, when calculating RMSDs between structures.

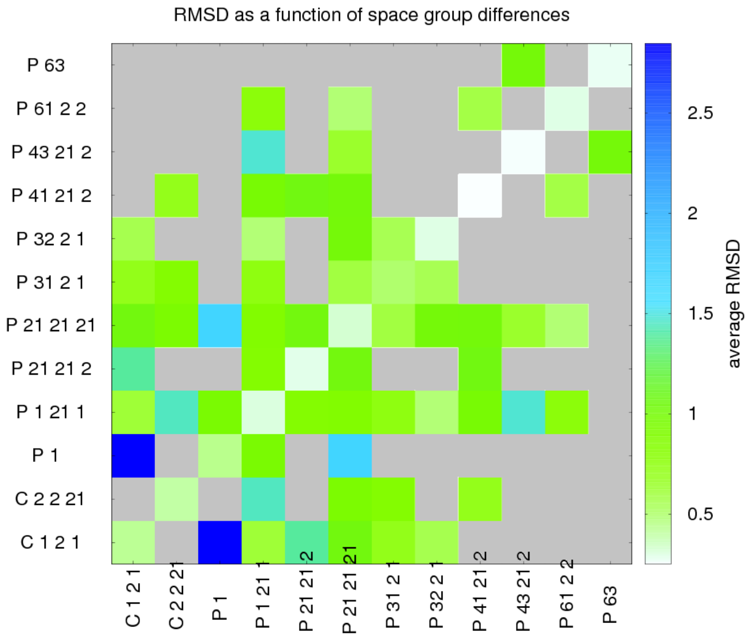


Figure 37 – Contour plot of average RMSD values for different combinations of space groups. The diagonal values in the middle correspond to cases of identical space groups. Off diagonal values are cases of different space groups. Only the most frequently occurring combinations of space groups (n >= 100) are plotted. Grey values correspond to infrequently occurring cases for which no average RMSD has been calculated (i.e. n < 100 for these combinations).

8.3.2.2 Ions

Because ions are often just present for buffer reasons, the presence of and the identity of the ions is often ignored. This is not an entirely correct procedure, as the RMSD can be shown to be dependent ($p < 0.01$) on whether the ions are strictly identical (<RMSD>$_{\text{identical}}$ = 0.28Å) or different (<RMSD>$_{\text{different}}$ = 0.36Å). However, this factor of identical or different ions can also be shown to have a cross effect ($p < 0.01$) with the factor of identical or different authors. The statistical RMSD difference is not solely due to different solutes.

If differences in ion concentration cause large structural changes, there is unfortunately no way to take into account ion concentration, as this information is never given in neither mmCIF nor PDB files. However, only a small conformational change is seen between two K^+ channel structures (PDB IDs 2hvj and 2hvk) solved at low and high K^+ ion concentrations.

8.3.3 Structural differences and discrete values

8.3.3.1 Number of chains

When plotting the RMSD as a function of the number of chains (appendix) for structure pairs that have identical space groups and zero mutations it can be seen that the RMSD varies with the number of chains. Therefore I choose to only focus on monomeric proteins from this point forward.

8.3.3.2 Number of mutations

As the number of mutations increases, I expect it to cause a larger structural difference between structures. I plot the RMSD between 41,941 monomeric structures from the same space group as a function of RMSD. When plotting the number of mutations against the RMSD a weak but statistically significant correlation (r = 0.23, $p_{r=0}$ < 0.01) is observed. This is a correlation between all data points. Many of those data points are for 0, 1 and 2 mutations. The data can be normalized by looking at mean RMSDs instead. If the correlation between the average RMSDs and the number of mutations is calculated instead, then the correlation coefficient, r_{averages}, increases to 0.97. Judging from the linear relationship, there seems to be no synergistic effect of mutations in general.

From Figure 38 it can be seen that comparison of structures with 2 amino acid differences occurs frequently. This is however not due to a large number of double mutants in the dataset. Instead it is a comparison of single point mutants against each other. Often mutants are solved by molecular replacement, and they will resemble the starting model. If this is the case, then two single point mutants solved by molecular replacement using the same starting model will resemble each other a lot. That is my explanation to why the average RMSD for cases of two amino acid differences is lower than that of a single and no amino acid differences. But at three sequence differences and beyond the average RMSD only increases (Figure 38).

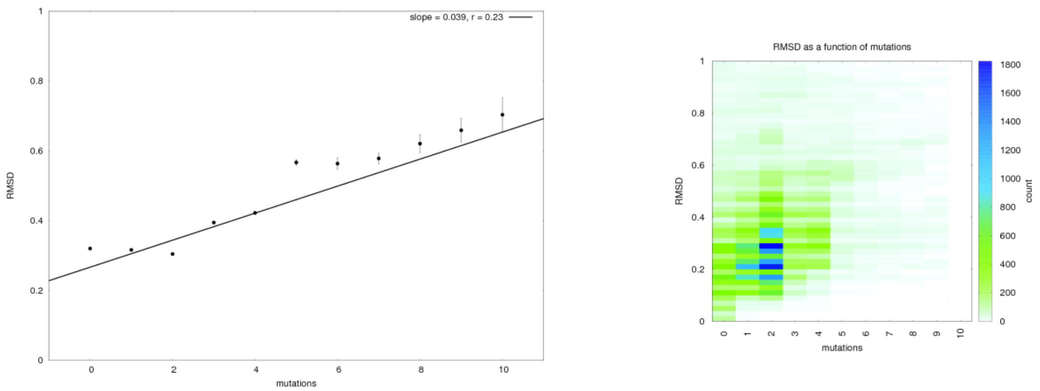


Figure 38 – Correlation between the number of mutations and the RMSD in the case of monomeric proteins crystallized in the same space group. n = 41941, r = 0.23, $t_{r=0}$ = 49.00, $p_{r=0}$ < 0.01, n_{averages} = 11, r_{averages} = 0.97, $p(r_{\text{averages}}$=0) < 0.01.

8.3.4 Structural differences and continuous values

After having presented the effect of nominal value parameters (e.g. space group) and discrete value parameters (number of mutations and chains) I turn to continuous value parameters. To rule out the effect of crystal contact differences, mutations and inter chain movements, I only analyze sequence identical monomeric proteins crystallized in the same space group.

Because pH, temperature, resolution and radius of gyration are continuous values, I cannot break the dataset down further relative to these values. Instead I perform an ordinary least squares regression of RMSD as a function of all four variables. Because the ionic strength along with the pH has an effect on the electrostatic interactions in proteins, I would like to have included it, but it is rarely reported. Instead I specifically look at pH difference, ΔpH, temperature difference, ΔT, the worst resolution of the two structures, max($d_{\text{min},1}$,$d_{\text{min},2}$), and the average radius of gyration, <r_G>. Only in 1,912 monomeric sequence identical structure pairs from the same space group are all four parameters known. I fit the parameters to a linear equation (equation 8-2) not considering any cross effects.

$$\text{RMSD} = \text{constant} + c_{\Delta\text{pH}}\Delta\text{pH} + c_{\Delta T}\Delta T + c_{d_{\min}}\max(\text{d}_{\text{min1}},\text{d}_{\text{min2}}) + c_{r_G} r_G \qquad \text{8-2}$$

	coefficient	std err	*t* statistic	*p* statistic
ΔpH	0.060	0.009	6.46	< 0.01
ΔT	-0.0013	0.0005	-2.58	< 0.01
max(d_{min})	0.12	0.02	7.94	< 0.01
<r_G>	0.0077	0.0017	4.49	< 0.01
constant	-0.045	0.030	-1.45	0.14

Table 11 – Results of the least squares regression. All factors have positive coefficients except the temperature difference. All coefficients are statistically significant, but statistical interactions can be present between the factors. d_{min} is the maximum resolution and r_g is the radius of gyration.

The result of the ordinary least squares regressions is shown in table Table 11. The most significant contribution to RMSD differences comes from a property not directly linked to the conformation of proteins. The variation in RMSD has the highest probability of being explained by the resolution (|t|=7.94). The conclusion from this observation is that low resolution structures (> 1.8Å) should be ignored, when analyzing the structural effect of mutations and ligands.

I expect a correlation between temperature difference and RMSD, because cryo structures are of better quality than room temperature structures. Surprisingly there is a negative correlation, when I do the ordinary least squares linear regression on all of the variables. That is because most cryo structures are solved at high resolution. Therefore a strong relationship exists between temperature and resolution and the *t* statistical value is larger for max(d_{min}) than it is for ΔT.

Next I look at the correlation between RMSD and individual continuous value parameters. When examining individual continuous value factors it is assumed that they are equally distributed over all protein pairs, so that any cross-effects can be ruled out. I start out with the temperature difference in order to try and explain the negative correlation achieved with the multivariable linear regression in this section.

8.3.4.1 Temperature differences

The average RMSD for cryo structures (anything below 273K) and other structures are shown in Table 12. It can be seen from the tabulated RMSD values that it is not whether a structure is cryo or not, which is decisive for its similarity to other structures. Instead it can be seen that structures solved by similar methods – and thus with a small temperature difference between them – have smaller RMSDs than a comparison between a cryo and non-cryo structure. Room temperature structures even have a slightly lower average RMSD than cryo structures. Maybe this can be explained by the fact that alternate conformations are provided in cryo structures, whereas it is almost always the average conformation, which is reported by standard temperature diffraction experiments. A comparison of two alternate conformations will yield a higher RMSD than the comparison of two averages.

	n	RMSD
Both structures cryo	1912	0.27
None of the structures cryo	644	0.24
Only one of the structures cryo	489	0.39

Table 12 – Average RMSD of cryo and non-cryo structures.

At higher temperatures a protein is thought to be more flexible. This should cause a lower precision in the determination of the atom positions from a more blurred electron density map. When the diffraction temperature difference is plotted against the RMSD (Figure 39), then it appears that the temperature difference makes a small contribution to the RMSD. The temperatures cluster at 100-140K and 273-298K. This is because diffraction experiments are either carried out at the standard temperature or at cryogenic conditions. Therefore the temperature differences are not evenly distributed across the whole range of the temperature scale. Therefore one should be careful about drawing any conclusions from the dependence of RMSD on temperature differences. However, it can be seen from Figure 39 that large temperature differences cause the largest RMSDs. This I do not necessarily attribute to the extent of the temperature difference as already discussed in this section. The reason large temperatures differences cause large RMSDs is because higher temperatures lead to conformational sampling. Instead of observing distinct conformations in the diffraction experiment, an average of conformations is observed. Multiple conformations will fit into the blurred electron density, and therefore large RMSDs are calculated as a consequence of a lack of precision in the positioning of atoms. Especially atoms with high B-factors that should not have been modeled will contribute to differences between high temperature structures. However cryo structures for which alternate locations of each atom are given will also lead to a high RMSD, if a low population state is compared to a high population state.

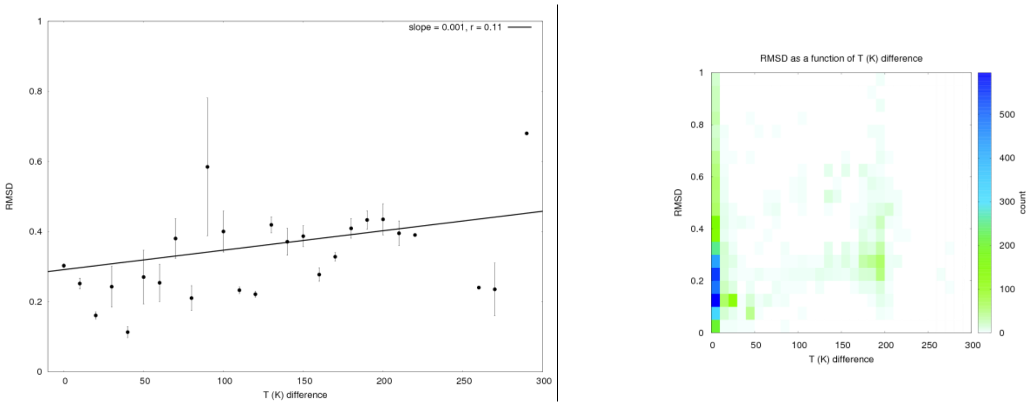


Figure 39 – A plot of the temperature difference between two X-ray structures and the RMSD between them. Larger temperature differences cause larger structural differences, but the relationship is not necessarily linear. ***n*** **= 4918,** ***r*** **= 0.09,** $t_{r=0}$ **= 6.98,** $p_{r=0}$ **< 0.01,** $r_{averages}$ **= 0.49,** $p_{r\_averages=0}$ **< 0.01.**

8.3.4.2 Resolution

The resolution of a protein structure is a measure of the atomic detail that can be read from the experimental data. The resolution is not a physiochemical property directly related to the conformation of a structure, but rather an indication of the error of the experimental determination of atom positions. In general proteins solved at cryo temperatures are solved at a better resolution (Rupp 2009) because conformational fluctuations are reduced and distinct conformations rather than average atom positions are observed. Distinct conformations can for example be seen in the recent ultra high 0.48Å atomic resolution cryo structure of crambin (3nir).(Schmidt, Teeter et al. 2011) Therefore if the resolution has an influence on the RMSD, then I also expect the temperature to have an influence on the RMSD. Higher temperatures will cause local flexibility and motion. This will in turn reduce the precision of the determination of atom positions. However it should also be mentioned, that cryo structures have been criticized for being locked into low populated conformations with perfect side chain packing due to a reduction of unit cell volume upon freezing; conformations which lie away from the path of functional motion.(Fraser, van den Bedem et al. 2011)

The resolution of the X-ray structure is expected to influence the RMSD, as the overall precision of the atom positions is dependent on the X-ray quality (Rupp 2009). Indeed, when plotting the average resolution of two structures against the RMSD (Figure 40) a weak but significant correlation (r = 0.16, $p_{r=0}$ < 0.01) is observed; i.e. poor resolution causes larger RMSDs. This is in correspondence with previous observations of the same phenomenon.(Carugo 2003) The above analysis was carried out with C_α atoms, which are less mobile than side chain atoms. Had the analysis been carried out with heavy atom RMSDs instead then the resolution would probably have had an even larger impact on the RMSD, since the conformation of side chains with high B-factors are heavily dependent on side chain rotamer libraries and choices of the modeler.

Integer value resolutions are overrepresented in the PDB. The majority of structures in the PDB have a resolution, d_{min}, of exactly 2.0 (Figure 40). This is an indication that the selection of a high resolution cutoff during data processing is skewed towards choosing inaccurate integer values. Indeed structures with a resolution reported at an accuracy of one tenth of an Å have shown to be of poorer quality (i.e. have higher clash scores) than those reported at one hundredth of an Å.(Read, Adams et al. 2011) If I exclude integer resolution values from my dataset of 6,254 values (2,551 at 1.0, 2,067 at 2.0, 741 at 3.0), then the correlation between RMSD and resolution increases from 0.16 to 0.31, which indicates a noise contribution from – by the author incorrectly chosen –integer resolution values.

Because of the expected relationship between temperature and resolution, I exclude non-cryo structures from the dataset and plot the RMSD as a function of resolution for cryo-structures only (Figure 41). The correlation coefficient increases from 0.16 to 0.29. The RMSD is somewhat constant for atomic resolution and high resolution (<1.8Å) structures, but rises thereafter as can be seen from the figure. The standard errors are also seen to increase above 1.8Å. And no structures with resolutions worse than 3Å are observed in this cryo data subset. An outlier is observed at 2.0Å. It could be due to incorrectly chosen high resolution cutoffs at integer values during data processing as already discussed. Low resolution structures rely on modeling software and side chain conformer libraries for high B-factor residues. Therefore a high RMSD is expected, if low resolution structures are not solved by the same person using the same procedure for the analysis of the diffraction pattern. Because the resolution has such a big influence on the final atom positions, one should - at least in the case of low resolution structures - rather compare electron densities to each other; instead of coordinates that are not always experimentally observed. To see an atom in the PDB with anything but 100% occupancy is quite rare, but the fact is that many atom positions are wholly or partly computationally determined; especially in the case of low resolution structures.

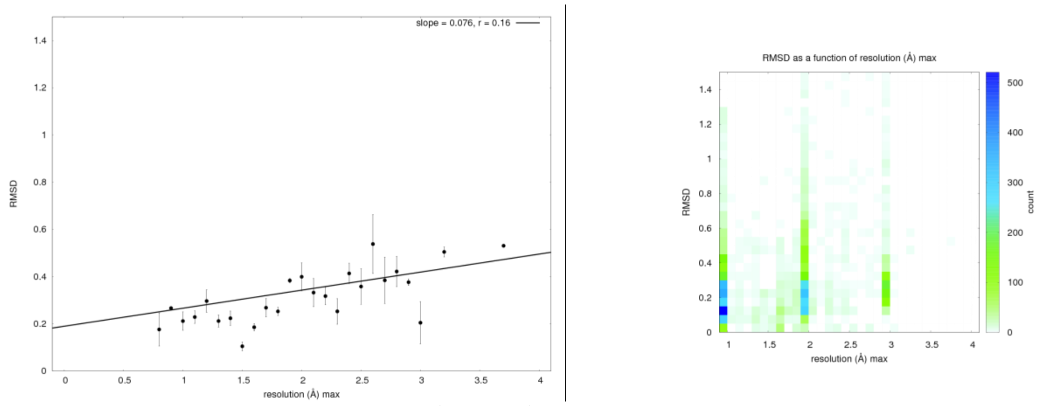


Figure 40 – The lowest resolution of two structures, max($d_{\mathrm{min},1}$, $d_{\mathrm{min},2}$) plotted against the RMSD. High resolution structures cause lower RMSDs. n = 6404, r = 0.16, r^2 = 0.03, slope = 0.08, $t_{r=0}$ = 13.23, $p_{r=0}$ < 0.01, r_{averages} = 0.64.

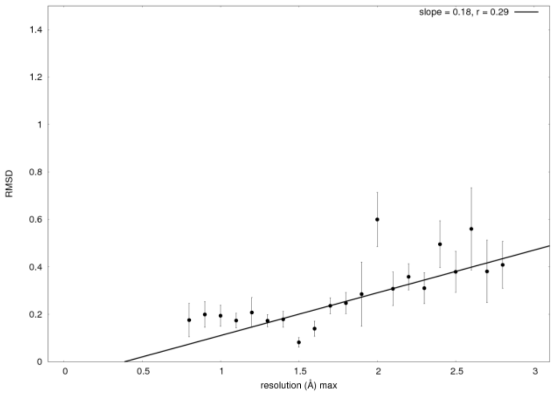


Figure 41 – The poorest resolution of two cryo structures plotted against the RMSD.

8.3.4.3 Difference in level of hydration

The Matthews coefficient, V_m, is a measure of the level of hydration in the unit cell. It is the volume of the unit cell divided by the molecular weight of the polymers in the unit cell(Matthews 1968); a reverse molecule density. Depending on the unit cell volume the crystal contacts will differ. In a large unit cell the motion of the surface exposed residues will not be restricted, and they will be free to move leading to high B-factors and a poor resolution. Furthermore water is disordered and scatters X-rays randomly. This could also contribute to a poor resolution. Therefore I expect large differences in the Matthews coefficient to lead to large RMSDs in some cases.

An extreme example of this is the structure pair 3bzc (V_m = 3.16) and 3bzk (V_m = 2.44). The protein has multiple domains. The two structures are sequence identical monomers from the same space group. Their only difference is their unit cell dimensions. The difference in unit cell volume leads to entirely different crystal forms (C_α RMSD 4.67Å) despite all other reported parameters being identical. The differences between 3bzc and 3bzk - that were both crystallized by vapor diffusion and the hanging drop method - are shown in Table 13. I find it most likely that the different uses of crystallization compounds caused the different crystal forms. It is a good example that a great number of variables need to be taken into consideration when analyzing causes of structural differences and it also shows that crystallography is more of an art than science. Unfortunately information about the crystallization buffer is not always provided in the PDB/mmCIF structure file. And when it is present, then it is usually in a non-standardized text format, which is challenging to parse.

	3bzc / crystal form I	3bzk / crystal form II
Crystal growth compounds	19% (w/v) PEG 3350 100mM bis-tris methane (pH 5.5) 0.17M ammonium sulfate 10% (v/v) glycerol	18% (w/v) PEG 4000 100mM sodium acetate (pH 4.6)
$\mathrm{pH}_{\mathrm{crystallization}}$	5.5	4.6
$T_{\mathrm{crystallization}}$ (K)	286	298
$T_{\mathrm{diffraction}}$ (K)	100	100
unit cell length *a* (Å)	57	56
unit cell length *b* (Å)	132	107
unit cell length *c* (Å)	144	140
resolution, d_{min} (Å)	2.3	2.3
average isotropic B factory, $<B_{\mathrm{iso}}>$	66.7	27.4
V_{m} (Å^3/Da)	3.16	2.44

Table 13 - Selected properties of 3bzc and 3bzk. *a*,*b*,*c* are the unit cell lengths. d_{min} is the reported highest resolution. V_{m} is the Matthews coefficient; i.e. the volume of the unit cell divided by the molecular weight of the protein.

To get a more general perspective on the effect of unit cell volume differences I plot the RMSD between two structures as a function of the subtracted difference in Matthews coefficient between them (Figure 43). I calculate the correlation coefficient, *r*, to be 0.32. If I only consider the averages of ΔV_{M} bins (0.01 steps) with more than 100 occurrences, then the correlation coefficient, $r_{\mathrm{averages},n>100}$, increases to 0.92.

However, the Matthews coefficient has been shown to be dependent on resolution.(Kantardjieff and Rupp 2003) Because the resolution has such a significant effect on the RMSD as shown in the previous section (8.3.4.2) I try to adjust the Matthews coefficient for the resolution. I calculate the expected Matthews coefficient at a 2.0Å resolution by linear extrapolation by doing a linear regression of V_{M} as a function of d_{min} (Figure 44), whereby I determine the slope of the fitted line to be 0.58. I calculate the correlation coefficient, *r*, between the adjusted ΔV_{M} and the RMSD to be 0.25; down from 0.32. The slope is reduced from 0.686 to 0.313. Although the correlation difference is not significant I take it as an indication, that part of the dependence of the RMSD on the V_{M} difference can be attributed to d_{min}. I also notice that the variance at each bin of V_{M} difference is reduced (Figure 45), which is further support to the argument of statistical interaction between V_{M} and d_{min}. The contribution from d_{min} to V_{M} is reduced by doing the adjustment.

I have also tried to include the Matthews coefficient as a variable in the ordinary least squares regression of RMSD as a function of all the continuous value variables. When resolution is left out of the equation, then RMSD is dependent on V_{M} ($t = 2.52$, $p = 0.012$). When both resolution and Matthews coefficient is included in the fitting procedure, then RMSD is not found to be significantly dependent on V_{M} ($p = 0.589$), but it is still dependent on resolution ($t = 7.94$, $p < 0.01$, Table 11). This is a further testament to the strong relationship and statistical interaction between the resolution and the Matthews coefficient. In terms of cause and effect it is a large Matthews coefficient, which causes a poor resolution.

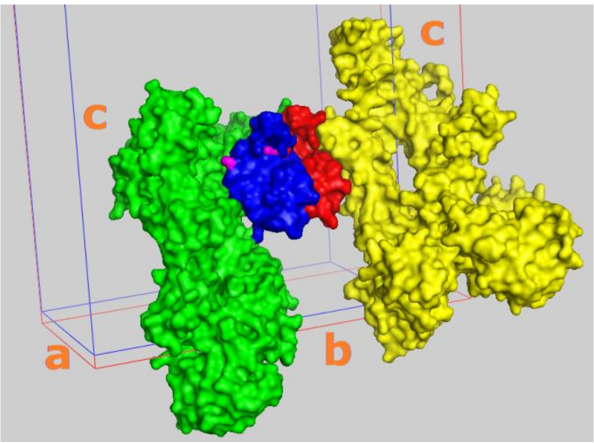


Figure 42 – Multi domain protein structures 3bzc and 3bzk aligned by residues 1-653. Residues 1-653 of 3bzc are shown in green. The SH1 domain (residues 654-730) is shown in blue (3bzc) and red (3bzk) respectively. One of the crystal contacts of 3bzc is shown in yellow. The unit cell axes are labelled with orange letters. It is the b-axis, which is 25Å shorter in 3bzk. Upon contraction of the b-axis the SH1 domain is forced to relocate because of the crystal contact shown in yellow. Residues 731-785 are not modelled because there is no discernable electron density due to a high degree of flexibility; the last visible C-terminal residue (730) is shown in magenta. Figure created with PyMol.

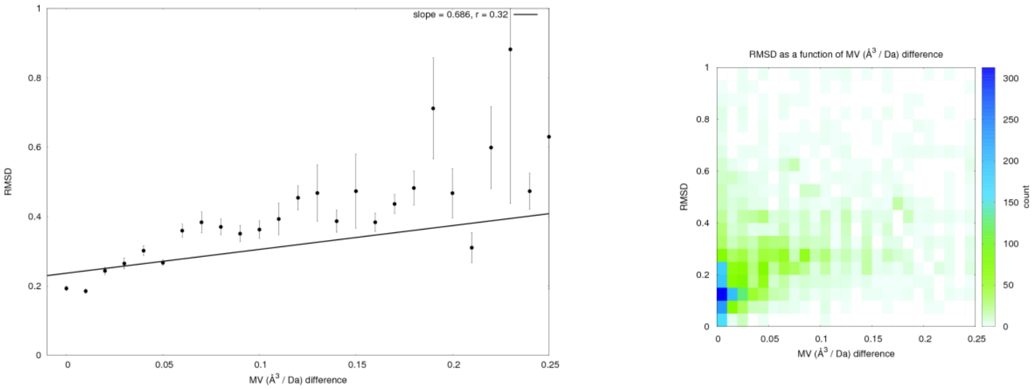


Figure 43 – RMSDs as a function of differences in crystal contacts as measured by Matthews coefficient differences. *n* = 4572, *r* = 0.32, r^2 = 0.10, slope = 0.69, $t_{r=0}$ = 22.70, $p_{r=0}$ < 0.01, $r_{\text{averages},n>100}$ = 0.92.

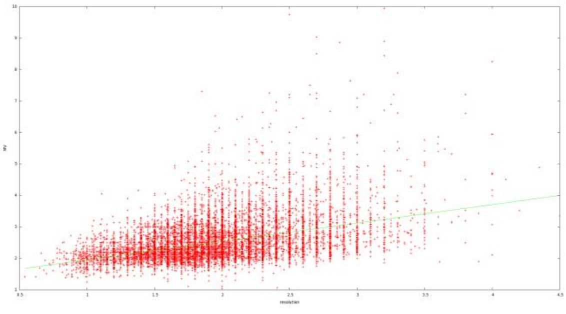


Figure 44 – The Matthews coefficient, V_m, as a function of resolution, d_{min}. The correlation coefficient is 0.43, the slope is 0.58 and the starting value at the y-axis interaction is 1.36. An empirical equation for the Matthews coefficient as function of resolution has been derived from V_m distributions at individual resolution bins/ranges(Kantardjieff and Rupp 2003), but here I choose to do a regression of all data points at all resolutions in a linear fashion.

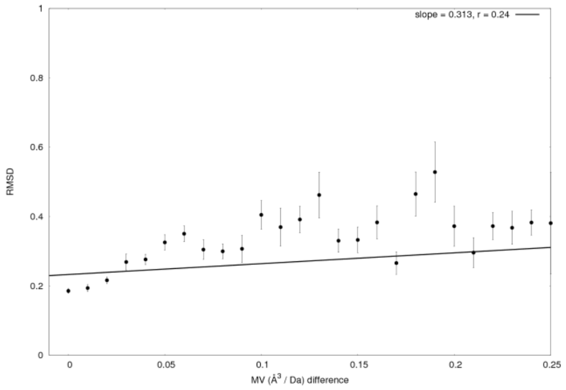


Figure 45 – Correlation between RMSD and Matthews coefficient adjusted for resolution. n = 4572, r = 0.25, r^2 = 0.06, slope = 0.32, $t_{r=0}$ = 17.23, $p_{r=0}$ < 0.01. Statistical calculations were performed on all data points, but not all data points are shown. Instead the average RMSD and the standard error at each bin are shown.

8.3.4.4 pH differences

At different pH values the protonation states of a protein will be different. The protein structure will change as a consequence of the change in electrostatic properties. pH induced conformational changes are well known and well documented.(Zand, Agrawal et al. 1971; Piszkiewicz 1974; Kukic, Farrell et al. 2009)

Reported pH values are associated with a large experimental error. In the lab the pH might have been measured in the buffer before initiation of protein crystallization. Sometimes the reported pH is just that of the buffering agent used. However, even if the pH is measured it changes during the hanging/sitting drop experiment, and pH values - and the difference between them - are therefore expected to be associated with a large experimental error.

I want to know if pH differences in general have an influence on protein conformations. The number of pH differences for sequence identical single chain proteins from the same space group totals 4686. I plot them against the RMSD in Figure 46. Most pH differences (3242) cluster at 0, which makes a linear regression a somewhat untrustworthy approach. It is however clear from Figure 46 that there is no immediate correlation between the RMSD and the pH difference between two structures. A minimum RMSD is not even observed at identical pH values. It should also be noted that many pH values cluster at integer values, because the exact pH in the protein crystal is not measured experimentally. The lack of correlation I partly attribute to the very inaccurately determined pH values. The ionic strength should also be considered, when analyzing the effect of electrostatic differences (pH differences) on protein structure. The ionic strength is however almost never reported in PDB/mmCIF structure files.

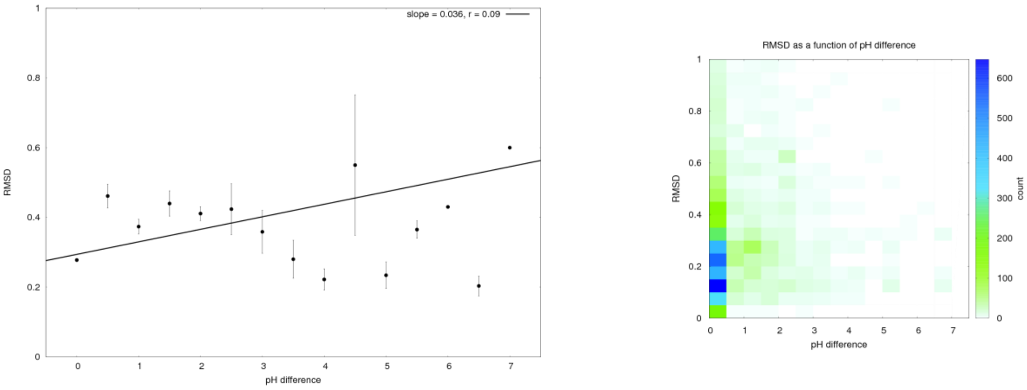


Figure 46 – A plot of the pH difference between two X-ray structures and the RMSD between them. Identical proteins and proteins mutated relative to each other are used for the analysis. Larger pH differences do not cause significantly larger structural differences (n = 4686, r = 0.09, r_{averages} = 0.04, $p(r_{\text{averages}}$=0) = 0.89).

8.3.4.5 Differences in protein size and shape

The length of the protein does not seem to be of importance to conformational difference between proteins as there is no correlation between RMSD and the number of residues (Figure 47). This is unexpected, as RMSD is often reported to be a size dependent measure of structural similarity. Specifically RMSD has been shown to be a dependent on the radius of gyration of a protein.(Cohen and Sternberg 1980; Reva, Finkelstein et al. 1998) However these results were achieved for random sequence structures upon analyzing the sequence similarity threshold at which a structure can be used for homology modeling. The structures in my dataset all have at least 95% sequence similarity and are not random sequence structures, which explain why I do not observe RMSDs to be size dependent. It seems that only in the case of random sequence structures is RMSD a size dependent property.

Another measure of protein size is the radius of gyration. It also reports on the shape of the protein. The radius of gyration is the square root of the mass weighted sum of squared distances between heavy atoms and the center of mass of the protein. Between two proteins with an identical number of residues a spherical/globular protein will have a smaller radius of gyration than an elongated protein. In the formula below n is the number of atoms in the protein and m_i and $\bar{x}_i$ is the mass and position of an atom, respectively.

$$r_{\mathrm{G}} = \sqrt{\sum_{i}^{n} m_i \left(\overline{x}_i - \sum_{i}^{n} m_i \overline{x}_i \right)^2} \tag{8-3}$$

I find that the RMSD is weakly correlated with the radius of gyration (r = 0.16, $p_{r=0}$ < 0.01, Figure 48). I do not find a correlation between the number of residues and the RMSD. Therefore I do not attribute the correlation between radius of gyration and RMSD to protein size, but rather protein shape. Elongated proteins with large radii of gyration have larger RMSDs upon movement than globular proteins, because of the larger distance between end points of the protein. For instance a conformational change in a elongated hinge bending protein will cause a larger RMSD the further away residues are located from the hinge point, and hinge motions are associated with the crossing of low energy barriers according to the theory of normal mode analysis, which makes it likely that the motion has a large amplitude. Therefore I attribute the correlation between RMSD and radius of gyration to protein shape and intrinsic dynamics rather than protein size.

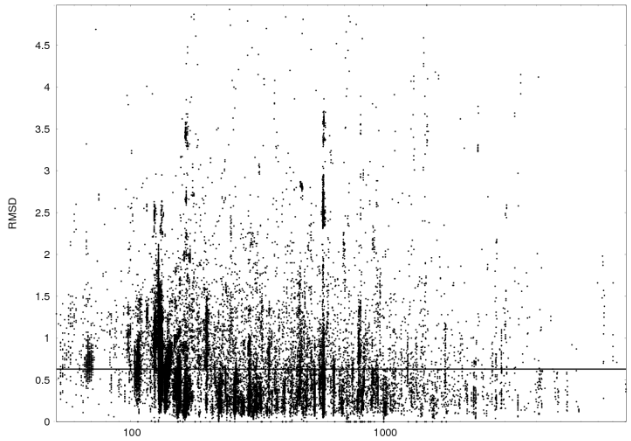


Figure 47 – There is no correlation between the number of residues in the protein pair being compared and the backbone RMSD between them. The axis on abscissa is logarithmic.

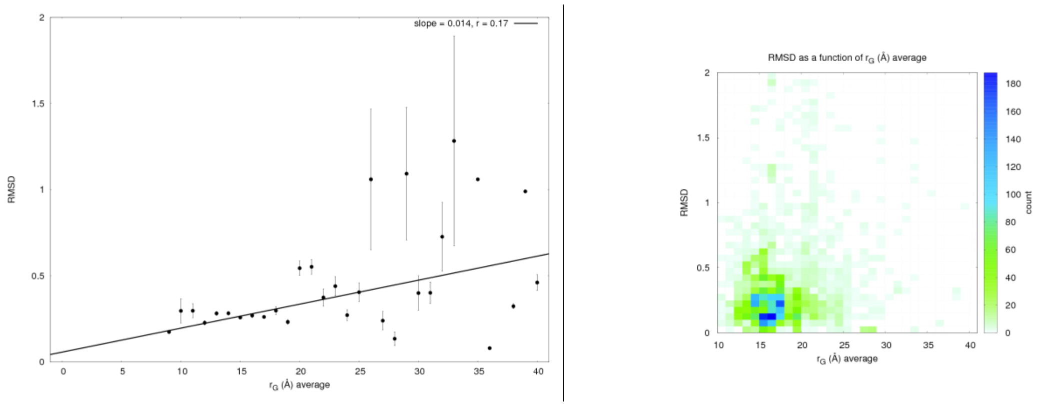


Figure 48 – RMSD as a function of radius of gyration. The regression is performed on all data points, but only the average and standard error at discrete values are shown. n = 4716, r = 0.16, r^2 = 0.03, slope = 0.013, $t_{r=0}$ = 11.26, $p_{r=0}$ < 0.01, r_{averages} = 0.48, p_{r_averages} < 0.01, $r_{\text{averages},n>100}$ = 0.69, p($r_{\text{averages},n>100}$ = 0) = 0.03.

8.3.5 Analysis of the structural effect of mutations

The main purpose of the investigation of the PDB as a whole was my interest in the structural effect of mutations. I wanted to know if the structural effect of a mutation propagates through the protein structure or the mutation only causes local structural changes if any at all. Now that it has been determined which properties contribute to structural differences, they can be accounted for and the structures of point mutants can be compared. Specifically I am interested in knowing about how far the structural effects of a mutation can travel in the protein. In this section I present the final results of my findings.

The four main parameters I use to quantify structural changes are all RMSD values. They are the RMSD of the position of C_α atoms in 3D space, the position of heavy atoms in 3D space, the position of Ramachandran angles on the 2D Ramachandran plot and the first side chain angle, χ_1. The coordinate RMSDs are dependent on the overall structure and can be quite large in the case of a large scale collective motion such as a hinge motion, whereas in theory the Ramachandran RMSD and the side chain dihedral RMSD are independent of remote structural changes.

I plot the average RMSD and the standard error at each distance from the site of mutation for each of my four selected variability parameters for a set of 11000 single chain mutant protein pairs in the PDB (section 8.3.5.2). This I do for 3 concentric spherical shells with their center at the site of the mutation and the radii [0:10[, [10:20[, [20:40[(Figure 49). However for any conclusions drawn from such plots to be valid I have to make sure that the RMSD values are normal distributed around the average RMSD. Therefore I plot the distribution of RMSDs for each of the 4 parameters at 4 selected distances (d=0, d=10, d=20, d=40) from the site of mutation (section 8.3.5.1). These 4 distances are chosen after an examination of the distribution of distances from the site of mutation (Figure 50). It is observed that the most frequent distance from the site of mutation is between d=10Å and d=20Å.

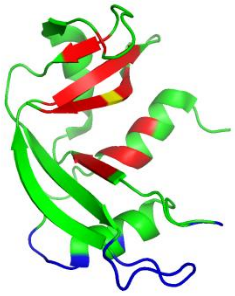


Figure 49 – An example of the 3 concentric spherical shells with radii [0:10[, [10:20[, [20:40[mapped onto a F120A mutant of ribonuclease A (PDB ID 1eic). The mutated residue is shown in yellow. The residues in the inner sphere are shown in red. The next sphere is green and the outer sphere is blue.

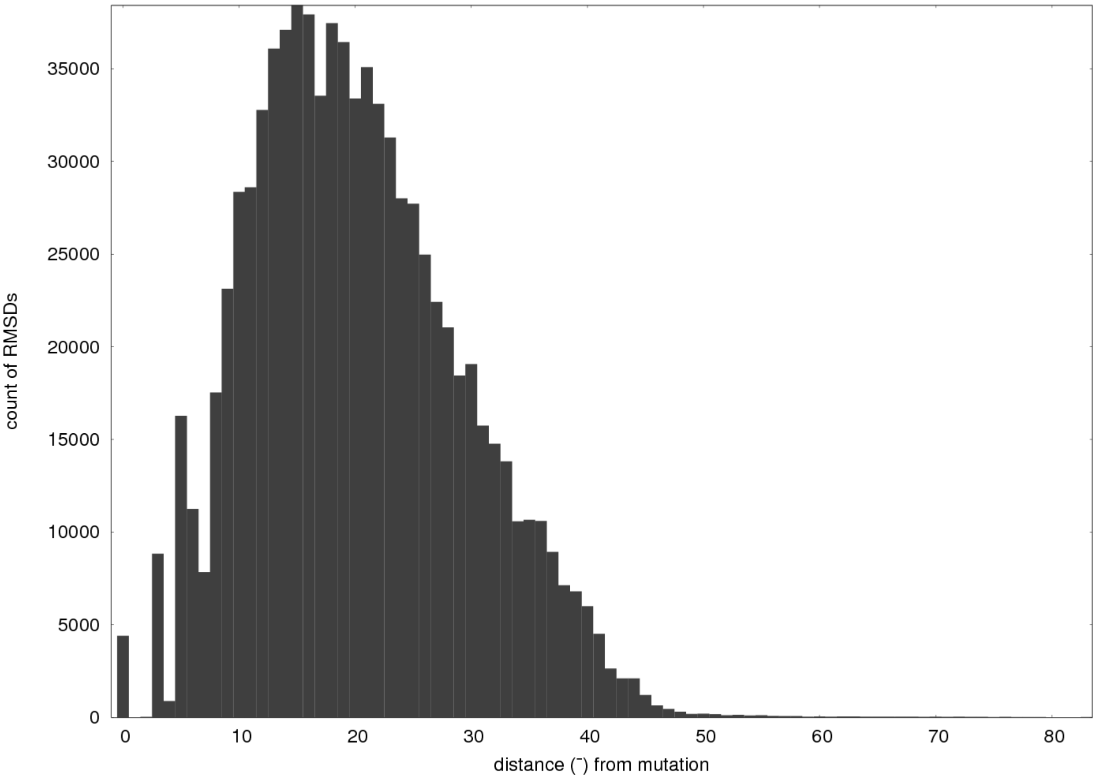


Figure 50 – Distribution of distances from the site of mutation. Most frequently residues are located 15Å from the site of mutation.

8.3.5.1 Distribution of RMSDs in spherical shells centered at the site of mutation

For each of three spherical shells centered at the site of mutation I plot the distribution of C_α RMSDs, heavy atom RMSDs, ϕψ RMSDs (°) and χ_1 RMSDs (Figure 51). I use 3 spherical shells (red, green and blue in Figure 51) to see, if the RMSD distributions shift as one moves further away from the site of mutation. To rule out the position of the mutation in the protein (for example at the center or at the end of an elongated protein) having an effect I compare the distributions 0Å from the site of mutation (Figure 51, left column) to those 20Å from the site of mutation (Figure 51, right column). If the proteins in the dataset display hinge motions - or other large scale concerted conformational changes - then residues far from the hinge point at the end of the protein will contribute more to the overall RMSD than residues in the vicinity of the hinge point.

The average heavy atom RMSD is higher than the average C_α RMSD, because side chain atoms in general have more degrees of freedom than C_α atoms; especially those at the surface not restricted by spatially neighboring residues in the interior of the protein. Because of the many degrees of freedom of surface residues the heavy atom RMSDs and χ_1 RMSDs are larger in the outer spherical shell.

Coordinate RMSDs depend on localized structural changes in hinge motion proteins. The Ramachandran angles are not affected by structural changes propagating through the structure the same way as overall coordinate RMSDs. ϕ,ψ RMSDs are therefore a more conservative measure of structural changes in the vicinity of the mutation.

χ_1 angles - unlike coordinates and to some extent ϕ,ψ angles - in general only attain selected angles (*trans*, *gauche*+, *gauche*-). Because of this the distribution of RMSDs are not smoothly distributed across the range of RMSDs. Instead a bimodal distribution is observed.

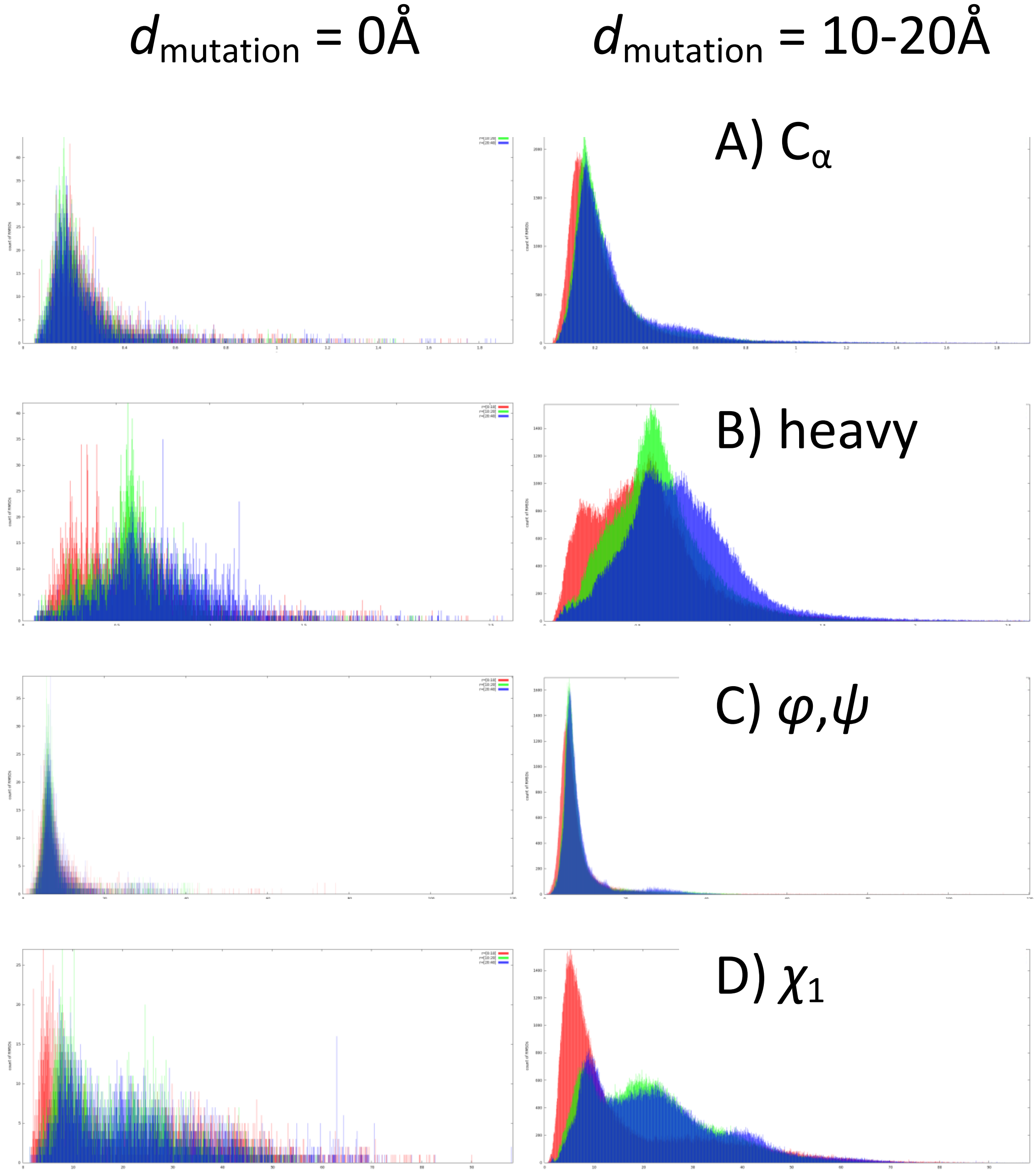


Figure 51 – Histograms of the distribution of A) C_α RMSDs (Å), B) heavy atom RMSDs (Å), C) $\varphi\psi$ RMSDs (°) and D) χ_1 RMSDs (°) within spherical shells with radii of 10Å (red), 20Å (Chargaff, Lipshitz et al.) and 40Å (blue) centered at the site of mutation (left column) and centered 10-20Å from the site of mutation (right column). A bimodal distribution is observed for χ_1 RMSDs, because χ_1 angles cluster at *trans*, *gauche*+ and *gauche*- conformations.

8.3.5.2 Average RMSDs as a function of distance from site of mutation

Having established in the previous section that RMSD values are normal distributed at various distances from the site of mutation I can proceed with the analysis looking just at averages at all integer value distances from the site of mutation.

For none of the four parameters do I observe the largest RMSDs close to the mutation (Figure 52). On the contrary the RMSD increases, as one moves further away from the site of mutation. If the increasing RMSD is due to spherical shells distant from the site of mutation containing many flexible surface residues, then the heavy atom and χ_1 RMSD should increase, while the C_α RMSD stays constant, because backbone atom positions are less susceptible to change due to surface exposure. However, the C_α RMSD also increases as one moves away from the site of mutation. In fact all four RMSDs increase; including the χ_1 RMSD, which is more a measure of surface exposure than overall conformation. Therefore I cannot simply attribute the rising RMSDs to the fact that residues distant from the site of mutation - and therefore also likely to be distant from the center of mass of the protein - are displaced further from the equilibrium position and the structure of comparison than residues at the center of the protein and close to the site of mutation. The rising RMSDs are a combination of both.

The conclusion is that, a single point mutation has limited structural effects in its near vicinity, and it has no effect on the experimentally observed overall conformation of a protein (Figure 38).

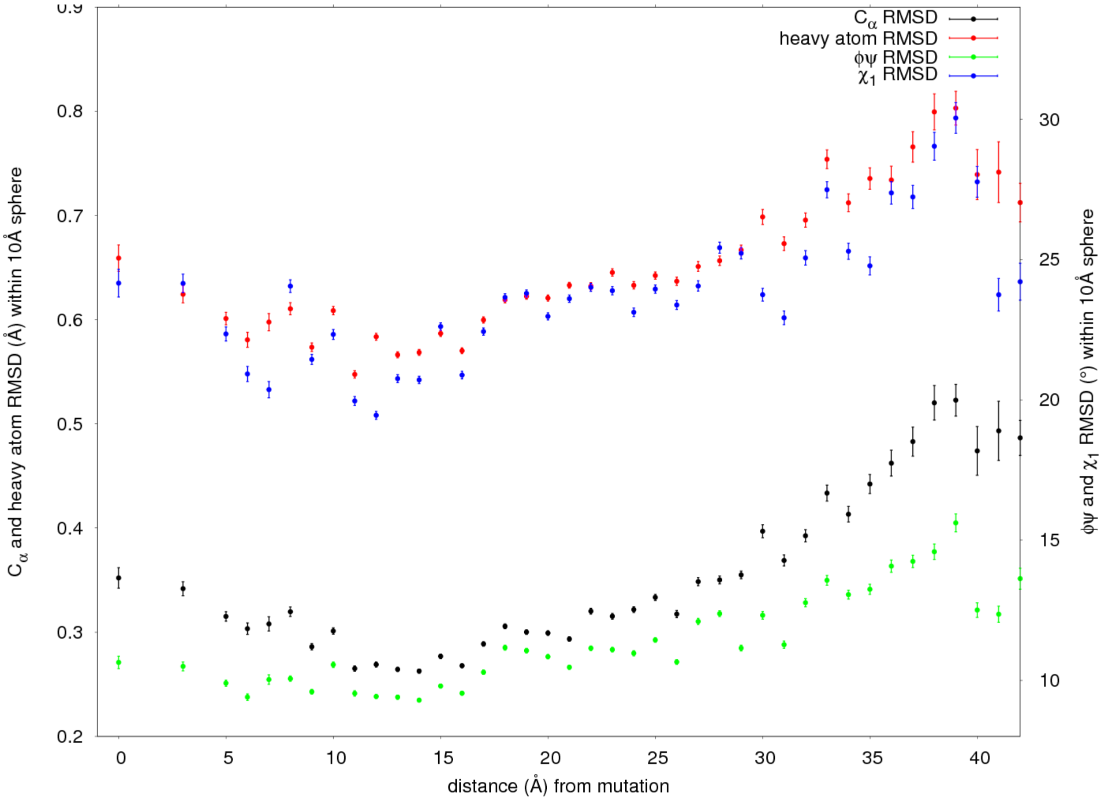


Figure 52 – Cα RMSD (Å, black), heavy atom RMSD (Å, red), ϕ/ψ RMSD (°, green) and χ_1 RMSD (°, blue) as a function of distance (Å) from the site of mutation.

8.4 Conclusions on the comparison of all structures in the PDB

I have shown that a single point mutation in general has no apparent effect on the overall conformation of a protein. Whereas conformational populations might be shifted for single point mutants, this effect does not translate to the reported X-ray structure. I have also shown that a single point mutation only has a limited structural effect in its nearby vicinity.

In fact the factors contributing the most to structural differences between protein structures are not mutations, but rather space group differences, differences in crystallographic resolution and differences in hydration levels. Different space groups and different hydration levels both change the crystal contacts, which is the explanation of their significant effect. Therefore - when comparing structures in order to identify the effect of a ligand or a mutation - those structures should be from the same space group as a very minimum. Although the number of mutations have an effect on the RMSD as the number of mutations increases, the small differences justifies using mutated structures for homology modeling and for solving the phasing problem of crystallography. When one takes into consideration that no structural changes – and no changes in the catalytic turnover rate – are observed upon mutating a stretch of 10 residues in T4 lysozyme to Alanine(Heinz, Baase et al. 1992), then it’s not a surprise that the effect of a single mutation is statistically insignificant.

With T4L and HEWL I observed that the author and starting model was of importance to structural similarity. For this large dataset of the whole PDB I observe the same. Despite the apparent similarity between experimental methods, there is an unexplained factor contributing to structural differences that differs between labs. This could be the starting model used for molecular replacement, buffering agents and ion concentrations not explicitly described in PDB files, modelling software and side chain conformer libraries and differences in the diffraction experiment after acquiring and harvesting protein crystals. What this factor is remains unexplained.

I observe a small effect of single point mutations and continuous value parameters on the RMSD. The largest non-discrete contribution to RMSD is from the resolution at which a structure was resolved. This is a parameter, which does not affect the true position of the atoms. Because resolution is not a physiochemical property, I conclude that observed structural differences are mostly attributable to structural errors caused by experimental limitations and conformational fluctuations at equilibrium.

Studies that report about conformational changes being caused by monatomic ions (e.g. 2hvj, 2hvk), electrons(Dai, Friemann et al. 2007), protons and other submolecular compounds are not wrong. But to quantify the small conformational change in these cases might not be justified considering the structural differences that are observed between identical proteins due to equilibrium fluctuations.

The main conclusions of this chapter are that single point mutations have no apparent effect on the overall conformation of proteins and the mutations only have an effect on residues in direct contact with the mutated residue. The effect of a mutation in general does not propagate throughout a structure.

9 Chapter 4 - Ligand binding by conformational selection: an analysis using elastic networks

9.1 Introduction

In this chapter I show that normal mode analysis (Nagel and Klinman) can be used to identify the ligand binding site of 4 proteins that have been shown to bind their ligand by conformational selection. The conformations of these proteins are not separated by an energy barrier, and NMA therefore describes the motion between the conformations. I also show that ligand binding residues in general display little conformational flexibility. I finally test my method of ligand binding site identification on a larger set of proteins, which have not been shown to bind their ligand by conformational selection.

9.1.1 The importance of ligand binding site prediction

Ligand binding has many functions and proteins can bind many different types of ligands. Enzymes can bind inhibitors and substrates(Blake, Koenig et al. 1965; Vocadlo, Davies et al. 2001), apo proteins can bind cofactors(Perutz, Rossmann et al. 1960), proteins can interact with other proteins (Vyas, Vyas et al. 1988; Stock, Mottonen et al. 1989; Milburn, Tong et al. 1990), antibody recognition(Padlan, Silverton et al. 1989; Stanfield, Fieser et al. 1990) and multimerization(Perutz, Rossmann et al. 1960)) and proteins can bind DNA (Aggarwal, Rodgers et al. 1988)). Structural genomics initiatives have over the past decade contributed to a large growth in protein structures with unknown function and unknown binding site.(Burley 2000) These binding sites cannot be found in existing databases based on sequential comparison to known structures.(Porter, Bartlett et al. 2004) The automated identification of ligand binding sites is therefore a relevant problem to focus on in order to identify natural binding sites and potential drug binding sites.

9.1.2 Characteristics of ligand binding sites

A characteristic of ligand binding sites is that they are fully or partly solvent exposed. Ligand binding sites are usually located at the largest cleft of the protein.(Laskowski 1996) The ligand binding pocket can be deep and narrow or shallow and wide. Protein-protein interactions often involve large intermolecular surfaces. Identification of these surfaces is difficult with geometric methods, which easily identify the largest pocket, but fail to identify large interacting surfaces. Even when a set of proposed interaction surfaces are available energetic methods sometimes fail to predict the correct interaction surfaces.(Henrick and Thornton 1998; Krissinel and Henrick 2007) A further challenge to ligand binding site prediction – especially for methods analyzing the micro environment of the protein – is the fact that ligand binding is associated with conformational changes of both the protein and the ligand(Burgen, Roberts et al. 1975; Birdsall, Feeney et al. 1980; LaPlante, Gillard et al. 2010). The conformational distribution of the protein is dependent on both the identity(Polgár and HalÁSz 1978; Brocklehurst, Willenbrock et al. 1983) and the concentration of the ligand. (Hammes, Chang et al. 2009)

9.1.3 Identification of ligand binding sites using normal mode analysis

In this work I identify ligand binding sites by utilizing normal mode analysis (Nagel and Klinman). If a ligand binds by conformational selection rather than induced fit, then there is no energy barrier between the free and ligand bound conformations, and the conformational change between the two conformations is a spontaneous thermal event, and NMA should be able to predict the amplitude and direction of the conformational change. In other words if conformational selection is dominant for a given protein, then it should be possible to predict the conformation of the protein in its ligand bound state using only information on the free state.(Seeliger and de Groot 2010; Meireles, Gur et al. 2011; Wako and Endo 2011) Perturbation of the elastic network by addition of nodes will change the amplitude and direction of calculated normal modes. Placing nodes at the ligand binding site is expected to cause the largest perturbation and thus permit the identification of the ligand binding site.

However, my NMA method is only expected to work in the case of ligand binding by conformational selection. In the case of ligand binding by induced fit, only in the presence of ligand will the conformational change across the energy barrier take place. NMA therefore in theory cannot be used to describe this motion and it cannot be used to identify the ligand binding site of proteins binding their ligand by induced fit. I identify the ligand binding sites of 4 model proteins known to bind their ligand by conformational selection, and I then examine 99 protein-ligand complexes where the structures of both the ligand free form and the ligand bound form are known. In the next section I briefly introduce conformational selection and the experimental results supporting ligand binding by conformational selection in 4 model proteins. In the following section I then introduce other geometry based methods of ligand binding site identification, that I benchmark my NMA method against.

9.1.4 Lock and key, induced fit and conformational selection

The ability of proteins to bind specific ligands has been studied for more than 100 years. Initially Emil Fischer as an explanation of enzyme specificity suggested that enzymes and their substrates bind in a lock and key mechanism.[a] (Fischer 1894) However, the lock and key mechanism fails to explain, why smaller analogous substrates are not catalyzed at saturation

[a] "Um ein Bild zu gebrauchen, will ich sagen, dass Enzym und Glucosid wie Schloss und Schlüssel zu einander passen müssen, um eine chemische Wirkung auf einander ausüben zu können."

levels, and therefore David Koshland introduced the induced fit theory, which hypothesized that reorganization of the active site upon ligand binding is a prerequisite for enzyme catalysis.(Koshland 1958; Yankeelov and Koshland 1965) The ability of proteins to populate several conformations is of high importance in enzyme-substrate binding(Zhou, Wlodek et al. 1998; Henzler-Wildman, Thai et al. 2007), protein-protein interactions, cooperative binding(Hill 1910) and allosteric regulation(Monod, Wyman et al. 1965; Popovych, Sun et al. 2006).

Opposed to the theory of induced fit is the theory of conformational selection caused by thermal motion. Conformational selection is in the literature also referred to as conformational selectivity, stabilization of conformational ensembles, population shift, selected fit, pre-existing equilibrium and fluctuation fit.(Vértessy and Orosz 2010) The difference between the theories of conformational selection and induced fit is shown in Figure 53. Whereas the theory of induced fit is a model that hypothesize that the binding of a ligand to a protein causes the conformational change of that protein(Koshland 1958), the theory of conformational selection is that the protein samples the ligand bound conformation independently of the presence of a ligand molecule. The sampling of the ligand free and bound conformations is a spontaneous/thermal event and the ability to sample the ligand bound conformation of the protein is an intrinsic property of the protein.

The four proteins, that I use for my analysis, which have been shown to bind their ligand by conformational selection, are shown in Table 14 and they are adenylate kinase (AdK)(Wolf-Watz, Thai et al. 2004; Henzler-Wildman, Lei et al. 2007; Henzler-Wildman, Thai et al. 2007), dihydrofolate reductase (DHFR)(Falzone, Wright et al. 1994; McElheny, Schnell et al. 2005; Boehr, McElheny et al. 2006; Bhabha, Lee et al. 2011), ribonuclease A (RNase A)(Rasmussen, Stock et al. 1992; Cole and Loria 2002; Beach, Cole et al. 2005; Kovrigin and Loria 2006) and peptidylprolyl isomerase A / cyclophilin A (CypA)(Eisenmesser, Millet et al. 2005; Fraser, Clarkson et al. 2009). For a more thorough walkthrough of the experimental evidence supporting conformational selection in my four model proteins I refer to the appendix.

In the case of AdK the loop covering the active site displays a large correlated motion when the enzyme is isolated in solution in the absence of substrate. It turns out that this motion is very similar to the conformational change observed upon ligand binding (Figure 54). Energy calculations also confirm that no barrier separates the ligand free and bound conformation of AdK.(Arora and Brooks 2007) Thus AdK presents a case where the conformational change necessary for function has been encoded in the protein structure, and the protein undergoes this conformational change, even when there is no substrate around. NMA should therefore be able to predict the motion between the open and closed conformation of AdK and other proteins binding their ligand by conformational selection.

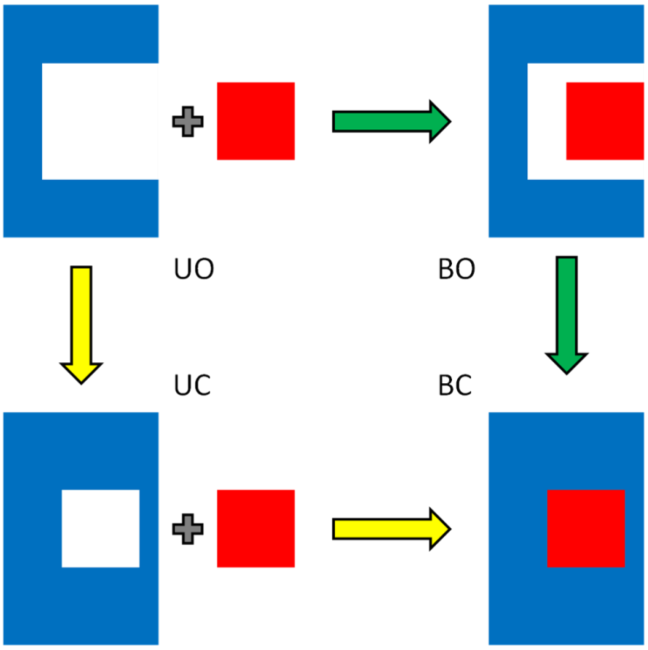


Figure 53 – A comparison of conformational sampling (yellow arrows) and induced fit (green arrows) upon binding of a ligand (red) by a protein (blue). The initial state in the top left corner is an unbound open (UO) conformation. The final state in the bottom right corner is a bound closed (BC) conformation. According to the induced fit theory the protein will follow the green path and change from open to closed conformation upon ligand binding. According to the theory of conformational sampling the protein will sample the closed conformation (bottom left) in the absence of the ligand by high frequency intrinsic motions; it is the closed conformation, which binds the ligand.

Adenylate kinase (AdK)
Dihydrofolate reductase (DHFR)
Ribonuclease A (RNase A)
Peptidylprolyl isomerase A / Cyclophilin A (CypA)

Table 14 – Proteins which have been shown to bind their ligand by conformational selection.

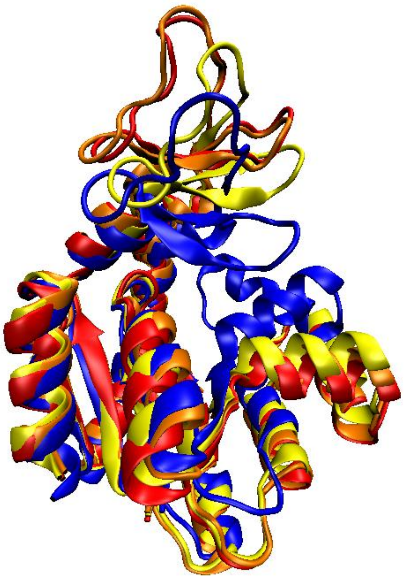


Figure 54 – Conformational variants of the enzyme adenylate kinase (AdK) in the absence of substrate (red, orange, yellow; PDB 2RH5) and in the presence of substrate (blue; PDB 2RGX). The two protein structures were solved in different space groups and the three chains in PDB structure 2RH5 all had different crystal contacts, which might have contributed to the conformational differences between the four chains. It would have been more optimal, had the structures been solved in the same space group containing only 1 chain, as effects of dissimilar crystal contacts could have been ruled out, but protein crystallization is rarely predictive.

9.1.5 Current methods of ligand binding site identification

Current successful methods for cavity finding are all structure based geometric methods. With the exception of ConCavity(Capra, Laskowski et al. 2009) and LIGSITE^{csc}(Huang and Schroder 2006) most of them assume that the largest pocket is the ligand binding site(Henrich, Salo-Ahen et al. 2009), because the active site of enzymes has been shown to often be located in the largest cleft of the protein.(Laskowski 1996) This highlights the importance of protein shape to protein function. Energetic methods for ligand binding site prediction also exist(Henrich, Salo-Ahen et al. 2009), but they require knowledge about the ligand, and they are further challenged by the fact that identical ligands have been shown to have very diverse ligand binding sites.(Kahraman, Morris et al. 2010)

ConCavity(Capra, Laskowski et al. 2009), which is based on the grid based method LIGSITE(Hendlich, Rippmann et al. 1997), makes use of sequential information for predicting, which of the identified pockets is the most likely ligand binding site. Because sequence conservation is highly predictive of identifying catalytic sites and ligand binding sites(Capra and Singh 2007), it is expected that ConCavity will perform well on the dataset. LIGSITE, which ConCavity is based on, is itself a grid based method, which in turn is based on POCKET(Levitt and Banaszak 1992), that identifies pockets by drawing lines in six different directions from each non-protein grid point. If a certain number of these oppositely directed lines pass the protein on both sides, then their originating grid point is classified as a pocket grid point.

POCASA(Yu, Zhou et al. 2010) is also a grid based method. The first step of POCASA however involves rolling a sphere on the surface of the protein to identify all pockets. Next all grid points in the pockets are categorized. For a grid point to be classified as a pocket grid point the number of nearby protein grid points has to exceed a certain number (18 by default) and the number of neighboring pocket grid points also has to exceed a given number (16 by default). This latter method is almost identical to how PASS(Brady and Stouten 2000) works, but POCASA performs better(Yu, Zhou et al. 2010) and is therefore used for benchmarking.

POCASA(Yu, Zhou et al. 2010) and ConCavity both predict the location and the shape of the binding pocket. However, here they are only benchmarked on their ability to predict the location of the bound ligand. Here the criterion for success is that the distance from the center of the ligand to the center of the predicted ligand binding site is smaller than a set threshold (6Å).

POCASA(Yu, Zhou et al. 2010) on a dataset of 48 structures – with the success criteria that the distance between any atom of the ligand and any point of the predicted binding site be less than 4Å – outperformed previous methods.(Laskowski 1995; Hendlich, Rippmann et al. 1997; Liang, Woodward et al. 1998; Brady and Stouten 2000; Huang and Schroder 2006; Weisel, Proschak et al. 2007) For a set of unbound and bound structures, POCASA had success rates of 75% and 77% respectively. Because POCASA is the best geometry based method, I use it for benchmarking here.

ConCavity does better than the previous methods it is built upon as measured by the area under the curve of ligand binding residue precision and recall for 332 proteins in the non-redundant LigASite (v7.0) dataset(Dessailly, Lensink et al. 2008), where precision is the ratio of true positives to true positives and false positives (residues that should have been selected, but were not) and recall is the ratio of true positives to true positives and false negatives (residues that should not have been selected, but were) (Figure 55). ConCavity is also expected to do better than POCASA, which otherwise outperforms the geometry based method LIGSITE, upon which ConCavity is based. The two have not been benchmarked against each other previously. ConCavity performs better than other methods on multi chain proteins. My dataset however is only made up of single chain proteins.

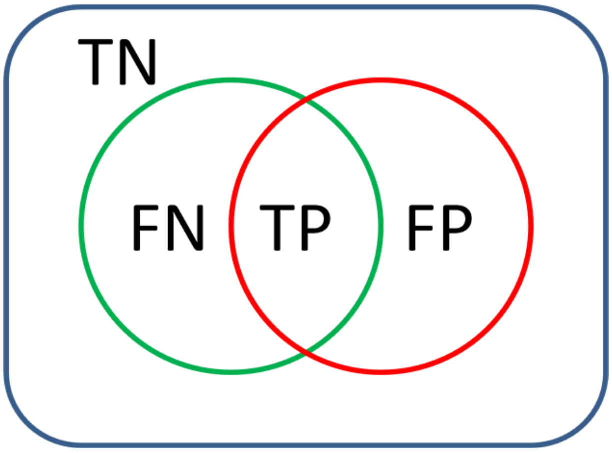


Figure 55 – Relationship between actual binding site (green circle), predicted binding site (red circle), true positives (TP), false positives (FP), false negatives (FN) and true negatives (Wang, Chumnarnsilpa et al.).

9.1.6 Ligand binding site residues are conserved and rigid

Residues in active sites are evolutionarily conserved(Capra and Singh 2007), and this fact is exploited by ConCavity to yield better results.(Capra, Laskowski et al. 2009) It has been speculated that flexible residues in active sites are more evolvable than rigid residues(Tokuriki and Tawfik 2009) and it has been shown that residues at protein-protein interaction surfaces have low B-factors(Neuvirth, Raz et al. 2004) and low calculated side-chain conformational entropy(Cole and Warwicker 2002), but no link between flexibility and evolvability has been established until now. In section 9.3.4 I show that residues with high conservation scores are never flexible as measured by B-factors. I show a correlation between the rigidity of resides and their inability to mutate over time. This establishes that conserved residues – such as those in active sites – are rigid, and it is the surrounding residues – rather than the active site residues – that are responsible for conformational changes observed upon ligand binding. I further show that the ligand contacting residues of my four model proteins are located at regions with low calculated flexibility. This gives the option of ruling out false positives (i.e. flexible residues), when using ligand binding site identification algorithms.

9.1.7 Structure of the results section of the chapter

In section 9.3.1 I present results showing that my NMA algorithm works as predicted. In section 9.3.2 I show that a few low energy modes contribute to the conformational change observed in the four model proteins binding their ligand by conformational selection. In section 9.3.2 I also show that my NMA algorithm correctly predicts the ligand binding site in the four model proteins. In section 9.3.3 I benchmark my NMA algorithm for ligand binding site identification against the already existing grid based methods. In section 9.3.4 I show that conserved residues are rigid as measured by their B-factors. I then show that ligand contacting residues in the four model proteins with NMA are calculated to be rigid. This latter result can be used to exclude false positive ligand binding sites identified by my NMA algorithm.

9.2 Methods

9.2.1 An introduction to normal mode analysis

A protein fluctuates among conformational sub-states at equilibrium. In the case of X-ray structures the equilibrium fluctuations are represented by B-factors and the X-ray structure is an average of the conformational sub-states. The B-factors describe the amplitude of the thermal motions of the individual atoms at equilibrium in the crystal. In the case of NMR structures the equilibrium fluctuations are represented by an ensemble of structures, which all satisfy the NOE restraints.(Lindorff-Larsen, Best et al. 2005) The ensemble of structures is generated from energy minimization of the conformations, which comply with the NMR restraints. So from both X-ray and NMR studies of protein structures it is known that folded proteins at equilibrium are not rigid structures but flexible structures.

One computational method for the study of protein dynamics is normal mode analysis. By definition, "normal mode analysis is the study of harmonic potential wells by analytic means."(Bahar and Cui 2005) A potential well is to be understood as a minimum of the potential energy at which the protein is at equilibrium. Normal mode analysis (Nagel and Klinman) can therefore be applied to the study of equilibrium fluctuations in particular. Interestingly the method in some cases also manages to simulate large-amplitude anharmonic fluctuations between protein configurations separated by a large energy barrier.(Zheng and Doniach 2003; Cui, Li et al. 2004; Li and Cui 2004; Tama, Feig et al. 2005; Zheng and Brooks 2005) Equilibrium fluctuations are of interest when studying for example protein stability. Functionally important motions are of interest when studying for example enzyme kinetics and thermodynamics. Protein folding of course also a functionally important motion, but the starting point of this motion is not a protein structure at equilibrium, which is what normal mode calculations are based on.

In section 9.2.3 the Gaussian network model - which is the basis of all my normal mode calculations - is presented. Using the Gaussian network model a macromolecule is represented as a network of springs. I will show that protein dynamics can be approximated by this elastic network model. In section 9.2.3 the mathematical theory of normal mode analysis and the calculation of normal modes are presented.

9.2.2 Application of normal mode analysis

Normal mode analysis has a wide range of applications. Because of its simplicity compared to MD simulations NMA has been used to study the dynamics of macromolecular structures on the Megadalton size scale.(Tama and Brooks Iii 2002; Tama, Wriggers et al. 2002; Tama, Valle et al. 2003) Computationally NMA is not limited by processing power like MD simulations, but rather by memory issues when diagonalizing the Hessian matrix. These can be circumvented by carrying out a coarse graining of the system.(Li and Cui 2002)

NMA has recently been utilized for exploring biologically relevant (i.e. non-linear) conformational transitions in proteins by running multiple steps of NMA in a “normal mode simulation” and doing geometric and force field structure correction between each step of NMA.(Ahmed, Rippmann et al. 2011) This method allows the study of non-linear motions of motions in proteins that are otherwise outside the size and/or time scale of what is feasible with MD simulations.

NMA can despite being a Harmonic method be used to describe non-thermal and energy barrier crossing motions, such as motions caused by ligand binding(Wako and Endo 2011) and changes in crystal hydration levels(Takayama and Nakasako 2011).

NMA can be used to determine protein structures from low resolution small angle X-ray scattering data(Miyashita, Gorba et al. 2011) and low resolution electron microscopy electron density maps(Tama, Miyashita et al. 2004; Tama, Miyashita et al. 2004) by finding alternative conformations of an already existing X-ray structure that are consistent with the low resolution experimental data.

NMA has been shown to correspond well with isotropic and anisotropic B-factors.(Kondrashov, Van Wynsberghe et al. 2007) I validate my algorithm by checking in section 9.3.1.1 that I get a correlation with B-factors. Because of the correlation with B-factors NMA could be used to determine, if a structure deposited in the PDB with falsified B-factors(Murthy, Smith et al. 2001; Ganesh, Muthuvel et al. 2005; Abdul Ajees, Gunasekaran et al. 2006; Ajees, Anantharamaiah et al. 2006) has realistic B-factors or not. Unfortunately nobody has developed such a check. Rather all existing checks are solely focused on close contacts and the geometry of the protein.

9.2.3 Theory of Normal Mode Analysis

Here the term "normal mode" is defined and its relevance to the potential energy of a molecule is outlined. The degrees of freedom of a molecule are the number of possible independent displacements of the molecule. An atom (e.g. Argon) in three dimensional space has 3 translational degrees of freedom. A two-atomic molecule (e.g. dinitrogen) has 3 translational degrees of freedom, 2 rotational degrees of freedom and 1 **vibrational degree of freedom**; i.e. the stretch. A nonlinear three-atomic molecule (e.g. water) has 3 translational degrees of freedom, 3 rotational degrees of freedom and 3 vibrational degrees of freedom; i.e. the symmetric and the antisymmetric stretch and the bend (**Figure 56**)(Atkins and de Paula 2002), p. 523)

A linear symmetric three-atomic molecule only has 2 rotational degrees of freedom, but 4 vibrational degrees of freedom; i.e. the symmetric and the antisymmetric stretch and two degenerate bends in the directions perpendicular to the length of the molecule and each other.

In general linear and nonlinear (e.g. a protein) polyatomic molecules have 3N-5 and 3N-6 vibrational degrees of freedom respectively. The vibrational degrees of freedom of a molecule are also termed the "**normal modes**" of the molecule (Atkins and de Paula 2002), p. 520).

In the case of water all three of the normal mode frequencies can be observed in an infrared spectrum, because all three normal mode vibrations cause a change in dipole moment (Atkins and de Paula 2002), p. 523).

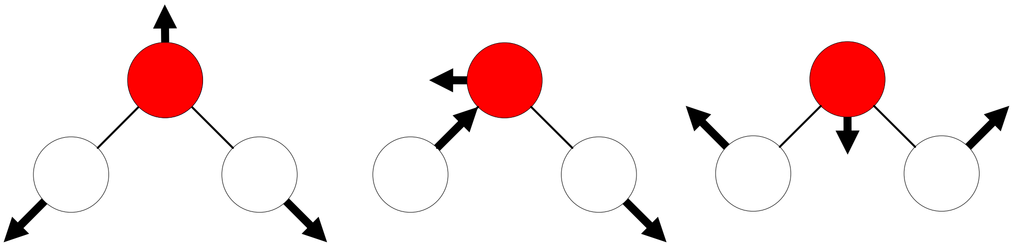


Figure 56 - The vibrational degrees of freedom of water; symmetric strecth, antisymmetric stretch and bend. Atom radii and vector lengths are not representative.

The set of all the coordinate combinations of the 3N coordinates of N atoms constitutes a 3N-dimensional **space** R^{3N}.[a] A set of 3N linearly independent[b] vectors in R^{3N} space forms a **basis**[c] for the R^{3N} space.[d]

The normal modes, translational modes and rotational modes are all **linearly independent**. A **superposition** of the normal modes yields the vibrational freedom/dynamics of the molecule.

The **potential energy** of a molecule is dependent on all 3N coordinates of a molecule with N atoms. The potential energy surface is therefore a function in the same R^{3N} space mentioned above. In vacuum the potential energy of a molecule is not dependent on translation and rotation. Therefore only normal modes that do not describe overall rotation and translation are of interest when calculating the change in potential energy of a molecule.

[a] "The n-dimensional space R^n is the set of all n-tuples $(x_1,x_2,x_3,...,x_n)$ of real numbers." Edwards, C. H. and D. E. Penney (1987). Elementary Linear Algebra., pp. 164

[b] Vectors are linearly independent if they are not linear combinations of each other.

[c] Vectors are a basis for a vector space if the vectors are linearly independent and the vectors span the vector space. Vectors span a vector space if every other vector in the vector space is a linear combination of the vectors.

[d] "Let V be an n-dimensional vector space and let S be a subset of V. Then if S is linearly independent and consists of n vectors, then S is a basis for V." Edwards, C. H. and D. E. Penney (1987). Elementary Linear Algebra., p. 184.

When a spring is compressed or extended, then the potential energy of the spring will increase. The normal mode associated with the smallest change in energy will therefore be one with the least compression and/or extension of the springs between atoms and vice versa. This can be illustrated with simple structures (Figure 57). In the case of these convex regular polyhedra the motion creating the largest strain on the springs between the atoms would be the motion towards or away from the centre. And this motion is exactly the one I calculate as the least favorable with my NMA algorithm (Figure 57).

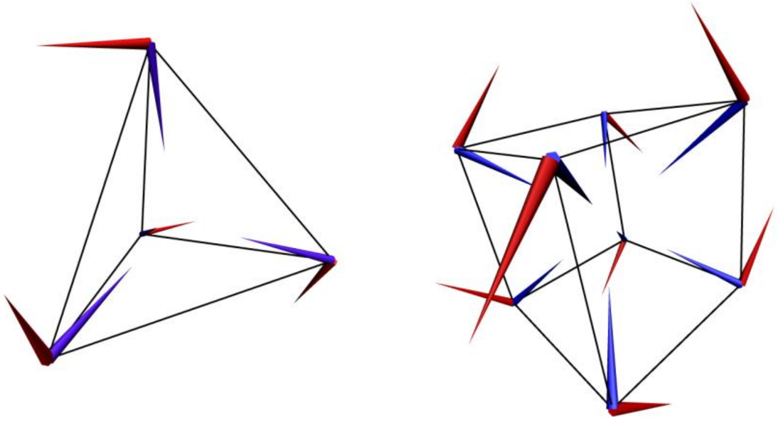


Figure 57 – Selected normal modes of the tetrahedron and the cube. The springs between nodes are shown in black. The most and the least favorable motion is shown in red and blue respectively. The blue eigenvectors all point towards the centre of the structures. Moving the atoms in this direction (or the opposite) would cause the greatest compression (or extension) of all the springs.

Sections 9.2.3.3 and 9.2.3.4 of this chapter will show that the normal modes can be calculated from a matrix of partial second derivatives of the potential energy at an energy minimum with respect to the coordinates. Sections 9.2.3.1 and 9.2.3.2 will further show how to calculate the elements of this matrix from the assumption of harmonic oscillations around the equilibrium. An overview of the theory sections in this chapter is shown below (Figure 58).

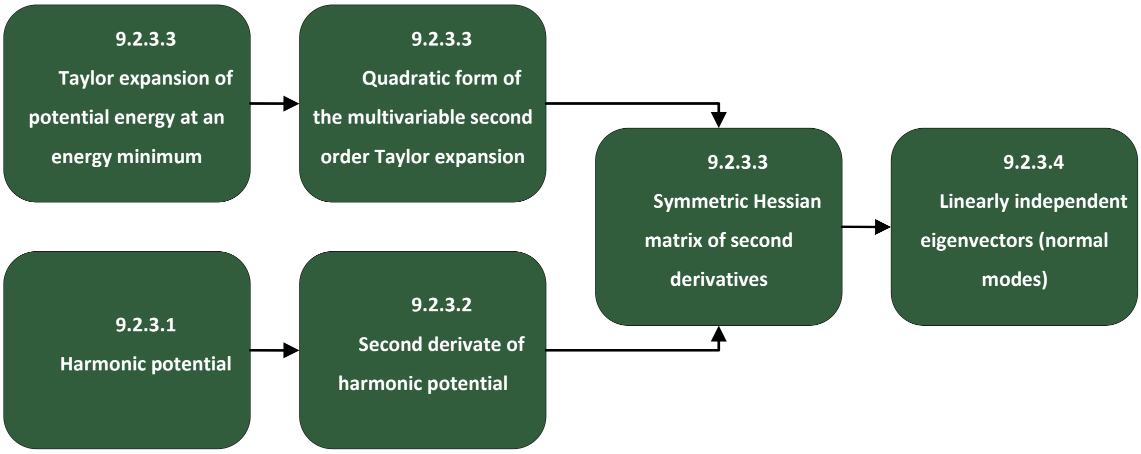

9.2.3.3
Taylor expansion of potential energy at an energy minimum
9.2.3.3
Quadratic form of the multivariable second order Taylor expansion
9.2.3.3
Symmetric Hessian matrix of second derivatives
9.2.3.4
Linearly independent eigenvectors (normal modes)
9.2.3.1
Harmonic potential
9.2.3.2
Second derivate of harmonic potential

Figure 58 – Overview of the theory sections in this chapter. Section numbers are given on top of the headlines.

9.2.3.1 Harmonic oscillators and the harmonic potential

Normal mode analysis is built on the principle of harmonic oscillations. An example of a harmonic oscillator can be seen in atoms and bonds. The two atoms represent two masses and the bond between them represents a spring. The atoms oscillate along the bond between them. This is an example of a harmonic oscillator, if the displacement of the masses from their equilibrium position follows Hooke's law. According to Hooke's law each mass experiences a force, F, proportional to the force constant, γ, of the spring and the displacement, s, from the equilibrium position, s^0.

$$F(s) = \gamma\left(s - s^0\right) \tag{9-1}$$

Integrating the force with respect to distance yields the change in potential energy upon displacement from equilibrium.

$$\begin{aligned} U(s) - U\left(s^0\right) &= \int F(s)ds \\ &= \int \gamma\left(s - s^0\right)ds \\ &= \tfrac{1}{2}\gamma\left(s - s^0\right)^2 \end{aligned} \tag{9-2}$$

The atoms in a fully folded protein can be assumed to perform harmonic oscillations around a conformational energy minimum. If the potential energy function is harmonic, then the potential energy, *U*, at equilibrium in three-dimensional space between atoms *i* and *j* connected by a single real spring (i.e. a bond) or a single phantom spring (cf. later discussion on Gaussian phantom networks) is

$$\begin{aligned} U\left(s_{ij}\right)-U\left(s_{ij}^{0}\right) &= ½\gamma_{ij}\left(s_{ij}-s_{ij}^{0}\right)^{2} \\ &= ½\gamma_{ij}\left(\left(\left(x_{j}-x_{i}\right)^{2}+\left(y_{j}-y_{i}\right)^{2}+\left(z_{j}-z_{i}\right)^{2}\right)^{½}-s_{ij}^{0}\right)^{2} \end{aligned} \qquad 9\text{-}3$$

The total potential energy, *U*, of a protein is equal to the sum of the potential energy between all of the interacting atoms.

$$\begin{aligned} U(s)-U\left(s^{0}\right) &= \sum_{i}^{N}\sum_{j}^{N}U\left(s_{ij}\right)-U\left(s_{ij}^{0}\right) \\ &= \sum_{i}^{N}\sum_{j}^{N}½\gamma_{ij}\left(s_{ij}-s_{ij}^{0}\right)^{2} \\ &= \sum_{i}^{N}\sum_{j}^{N}½\gamma_{ij}\left(\left(\left(x_{j}-x_{i}\right)^{2}+\left(y_{j}-y_{i}\right)^{2}+\left(z_{j}-z_{i}\right)^{2}\right)^{½}-s_{ij}^{0}\right)^{2} \end{aligned} \qquad 9\text{-}4$$

In the next section the second derivative of the harmonic potential is calculated. The second derivatives are elements of the Hessian matrix which will be introduced in section 9.2.3.3.

9.2.3.2 Second derivative of the Harmonic potential, calculating the elements of the Hessian matrix

By applying the power and the chain rule of differentiation the first derivative of the Hookean potential (eq. 9-4) with respect to the vector component *x* of the atom *i* is found to be

$$\begin{aligned} \frac{\partial U}{\partial x_{i}} &= \frac{\partial\left(\sum_{i}\sum_{j}½\gamma\left(\left(\left(x_{j}-x_{i}\right)^{2}+\left(y_{j}-y_{i}\right)^{2}+\left(z_{j}-z_{i}\right)^{2}\right)^{½}-s_{ij}^{0}\right)^{2}\right)}{\partial x_{i}} \\ &= ½\gamma\sum_{j}(-1)\left(2\left(x_{j}-x_{i}\right)\right)\left(½\left(s_{ij}{}^{2}\right)^{-½}\right)\left(2\left(\left(s_{ij}{}^{2}\right)^{½}-s_{ij}^{0}\right)\right) \\ &= -\gamma\sum_{j}\left(x_{j}-x_{i}\right)\left(1-s_{ij}^{0}s_{ij}{}^{-1}\right) \end{aligned} \qquad 9\text{-}5$$

Because $s_{ij}{}^{0}$ is the equilibrium position, all of the first derivatives equal zero at the energy minimum, $s_{ij}=s_{ij}{}^{0}$.

By applying the product rule to the first derivatives from above, the second derivatives at the energy minimum $s_{ij}=s_{ij}{}^{0}$ are found to be

$$
\begin{aligned}
\frac{\partial^2 U}{\partial x_i \partial y_j} &= \frac{\partial \frac{\partial U}{\partial x_i}}{\partial y_j} \\
&= \frac{\partial\left(-\gamma \sum_j \left(x_j - x_i\right)\left(1 - s_{ij}^0 s_{ij}^{-1}\right)\right)}{\partial y_j} \\
&= -\gamma\left(x_j - x_i\right)\left(2\left(y_j - y_i\right)\right)\left(½\left(s_{ij}^{\ 2}\right)^{-½}\right)\left(s_{ij}^0 s_{ij}^{-2}\right) \\
&= -\gamma\left(x_j - x_i\right)\left(y_j - y_i\right) s_{ij}^0 s_{ij}^{-3} \\
&= -\gamma\left(x_j - x_i\right)\left(y_j - y_i\right) s_{ij}^{-2}
\end{aligned}
\qquad 9\text{-}6
$$

and

$$
\begin{aligned}
\frac{\partial^2 U}{\partial x_i \partial x_j} &= \frac{\partial\left(-\gamma \sum_j \left(x_j - x_i\right)\left(1 - s_{ij}^0 s_{ij}^{-1}\right)\right)}{\partial x_j} \\
&= -\gamma\left(\left(1 - s_{ij}^0 s_{ij}^{-1}\right) + \left(x_j - x_i\right)\left(x_j - x_i\right) s_{ij}^0 s_{ij}^{-3}\right) \\
&= -\gamma\left(x_j - x_i\right)\left(x_j - x_i\right) s_{ij}^{-2}
\end{aligned}
\qquad 9\text{-}7
$$

It is these second derivatives that are the elements of the Hessian matrix, which will be introduced in the next section.

If the potential energy is a continuous function, then by the rule of mixed partial derivatives, the remaining second derivatives can be calculated using the equations below.

$$
\begin{aligned}
&\frac{\partial^2 U}{\partial x_i \partial y_j} \equiv \frac{\partial^2 U}{\partial y_j \partial x_i} = -\gamma\left(x_j - x_i\right)\left(y_j - y_i\right) s_{ij}^{-2} = -\gamma\left(y_j - y_i\right)\left(x_j - x_i\right) s_{ij}^{-2} = \frac{\partial^2 U}{\partial y_i \partial x_j} \equiv \frac{\partial^2 U}{\partial x_j \partial y_i} \\
&\frac{\partial^2 U}{\partial x_i \partial z_j} \equiv \frac{\partial^2 U}{\partial z_j \partial x_i} = -\gamma\left(x_j - x_i\right)\left(z_j - z_i\right) s_{ij}^{-2} == -\gamma\left(z_j - z_i\right)\left(x_j - x_i\right) s_{ij}^{-2} = \frac{\partial^2 U}{\partial z_i \partial x_j} \equiv \frac{\partial^2 U}{\partial x_j \partial z_i} \\
&\frac{\partial^2 U}{\partial y_i \partial z_j} \equiv \frac{\partial^2 U}{\partial z_j \partial y_i} = -\gamma\left(y_j - y_i\right)\left(z_j - z_i\right) s_{ij}^{-2} = -\gamma\left(z_j - z_i\right)\left(y_j - y_i\right) s_{ij}^{-2} = \frac{\partial^2 U}{\partial z_i \partial y_j} \equiv \frac{\partial^2 U}{\partial y_j \partial z_i} \\
&\frac{\partial^2 U}{\partial x_i \partial x_j} \equiv \frac{\partial^2 U}{\partial x_j \partial x_i} \\
&\frac{\partial^2 U}{\partial y_i \partial y_j} \equiv \frac{\partial^2 U}{\partial y_j \partial y_i} \\
&\frac{\partial^2 U}{\partial z_i \partial z_j} \equiv \frac{\partial^2 U}{\partial z_j \partial z_i}
\end{aligned}
\qquad 9\text{-}8
$$

The second derivatives - with respect to components of atom *i* only - are the negative sum of the second derivates - with respect to components of atom *i* and *j* - as shown below. Put differently, the sum of the second derivatives of the potential energy, which is the sum of forces, equals zero for each atom.

$$
\begin{aligned}
\frac{\partial^2 U}{\partial x_i \partial y_i} &= \frac{\partial\left(-\gamma \sum_j \left(x_j - x_i\right)\left(1 - s_{ij}^0 {s_{ij}}^{-1}\right)\right)}{\partial y_i} \\
&= -\gamma \sum_j \left(x_j - x_i\right)\left(-1\right)\left(2\left(y_j - y_i\right)\right)\left(½\left({s_{ij}}^2\right)^{-½}\right)\left(s_{ij}^0 {s_{ij}}^{-2}\right) \\
&= \gamma \sum_j \left(x_j - x_i\right)\left(y_j - y_i\right) s_{ij}^0 {s_{ij}}^{-3} \\
&= -\sum_j \frac{\partial^2 U}{\partial x_i \partial y_j}
\end{aligned}
\qquad 9\text{-}9
$$

And

$$
\begin{aligned}
\frac{\partial^2 U}{\partial x_i \partial x_i} &= \sum_j -\frac{\partial U}{\partial x_i}\left(x_j - x_i\right)^{-1} + \gamma\left(x_j - x_i\right)\left(x_j - x_i\right) s_{ij}^0 {s_{ij}}^{-3} \\
&= \gamma \sum_j \left(x_j - x_i\right)\left(x_j - x_i\right) {s_{ij}}^{-2} \\
&= -\sum_j \frac{\partial^2 U}{\partial x_i \partial x_j}
\end{aligned}
\qquad 9\text{-}10
$$

In the next section it will be shown how the second derivatives of the Harmonic potential calculated in this section constitute the Hessian matrix.

9.2.3.3 Taylor expansion of the potential energy function, a quadratic approximation

This section will that the potential energy can be approximated by a quadratic function, which can be written on matrix form. The elements of this matrix were calculated in the previous section.

The potential energy function, *U*, of a protein is a function of the structure and thus of the Cartesian coordinates of the *N* atoms. This yields a total of 3*N* variables for the potential energy function.

$$
U\left(x_1, y_1, z_1, \dots, x_n, y_n, z_n\right) = U\left(\boldsymbol{s}\right) \qquad 9\text{-}11
$$

where ***s*** is the vector representing the coordinates.

The potential energy can be approximated using a Taylor series. The *K*th order Taylor expansion around a conformational energy minimum, $s = s^0$, of the potential energy function, *U*, is

$$\begin{aligned} U(\boldsymbol{s}) &= \sum_{k=0}^{K} \frac{1}{k!}\left(\sum_{n}^{3N}\left(s_n - s_n^0\right)\frac{\partial}{\partial s_n}\right)^k U\left(\boldsymbol{s}^0\right) \\ &= \left(1 + \left(\sum_{n}^{3N}\left(s_n - s_n^0\right)\frac{\partial}{\partial s_n}\right) + ½\left(\sum_{n}^{3N}\left(s_n - s_n^0\right)\frac{\partial}{\partial s_n}\right)^2 + \sum_{k=3}^{K}\frac{1}{k!}\left(\sum_{n}^{3N}\left(s_n - s_n^0\right)\frac{\partial}{\partial s_n}\right)^k\right)U\left(\boldsymbol{s}^0\right) \end{aligned} \qquad \text{9-12}$$

If $s = s^0$ is a minimum of the function, *U*, then the first derivatives $\frac{\partial U\left(s_n^0\right)}{\partial s_n} = 0$ equal zero. If third order and higher terms are neglected, then it yields a symmetric **quadratic approximation(Lifson and Warshel 1968; Levitt, Sander et al. 1985)** for the potential energy function at the local energy minimum $s = s^0$.

$$U(\boldsymbol{s}) - U\left(\boldsymbol{s}^0\right) = ½\left(\sum_{n=1}^{3N}\left(s_n - s_n^0\right)\frac{\partial}{\partial s_n}\right)^2 U\left(\boldsymbol{s}^0\right) \qquad \text{9-13}$$

The square of a homogenous polynomial[a] of first degree in 3*N* variables is obviously a homogeneous polynomial of second degree in 3*N* variables.

$$\begin{aligned} U(\boldsymbol{s}) - U\left(\boldsymbol{s}^0\right) &= ½\left(\sum_{n=1}^{3N}\left(s_n - s_n^0\right)\frac{\partial}{\partial s_n}\right)^2 U\left(\boldsymbol{s}^0\right) \\ &= ½\left(\sum_{i=1}^{3N}\left(\left(s_i - s_i^0\right)\frac{\partial}{\partial s_i}\sum_{j=1}^{3N}\left(s_j - s_j^0\right)\frac{\partial}{\partial s_j}\right)\right)U\left(\boldsymbol{s}^0\right) \\ &= ½\left(\sum_{i=1}^{3N}\sum_{j=1}^{3N}\left(\left(s_i - s_i^0\right)\left(s_j - s_j^0\right)\frac{\partial^2}{\partial s_i \partial s_j}\right)\right)U\left(\boldsymbol{s}^0\right) \end{aligned} \qquad \text{9-14}$$

A homogeneous polynomial of degree 2 is also a **quadratic form**.(Edwards and Penney 1987)

$$U(\boldsymbol{s}) - U\left(\boldsymbol{s}^0\right) = ½\boldsymbol{s}^T\boldsymbol{Hs} \qquad \text{9-15}$$

s is the 3*N*x1 column vector with the elements $\left(s_n - s_n^0\right)$ and ***s***^{T} is the transpose; i.e. the 1x3*N* row vector. ***H*** is the matrix for the quadratic form. ***H*** is a 3*N*x3*N* matrix with the second derivative elements $\frac{\partial^2 U\left(\boldsymbol{s}^0\right)}{\partial s_i \partial s_j}$ (eq. 9-8). A matrix with second derivatives is named a Hessian matrix. If the potential energy function is continuous, then by the rule of mixed partial derivatives

[a] A homogeneous polynomial is a polynomial with all terms having the same degree. All symmetric polynomials, which are polynomials that are invariant upon permutation of the variables of the polynomial, are thus homogeneous polynomials.

$$\frac{\partial^2 U\left(\boldsymbol{s}^0\right)}{\partial s_i \partial s_j} = \frac{\partial^2 U\left(\boldsymbol{s}^0\right)}{\partial s_j \partial s_i} \qquad 9\text{-}16$$

Hessian matrices are thus symmetric.

The 3*N*x1 **r**-vectors are elements of the R^{3N} space of the potential energy function of 3N variables. 3*N* linearly independent[a] ***r***-vectors form a basis for the R^{3N} space. In the next section I show that a set of 3*N* linearly independent ***r***-vectors, which span the R^{3N} space, are the eigenvectors of the symmetric Hessian matrix.

9.2.3.4 Diagonalization of the Hessian matrix

This section will show how to derive the normal modes from the Hessian matrix.

The *n* x *n* square matrices ***A*** and ***D*** are called similar if there exists an invertible matrix ***P*** such that

$$\boldsymbol{D} = \boldsymbol{P}^{-1}\boldsymbol{AP} \qquad 9\text{-}17$$

A square matrix is called diagonalizable (Edwards and Penney 1987), p.273) if it is similar to a diagonal matrix in which all off-diagonal elements are zero. If ***A*** is diagonalizable, then ***D*** is an eigenvalue matrix[b] and ***P*** is an eigenvector matrix with *n* **linearly independent eigenvectors**.[c] If a square matrix is diagonalizable and the **eigenvector matrix is orthogonal**[d], then a square matrix is called orthogonally diagonalizable (Edwards and Penney 1987), p.293). According to the spectral theorem of linear algebra[e] the Hessian matrix, ***H***, is orthogonally diagonalizable. Thus the quadratic form for a set of linearly independent **r**-vectors can be written.

$$\tfrac{1}{2}\boldsymbol{R}^{T}\boldsymbol{HR} = \tfrac{1}{2}\boldsymbol{R}^{-1}\boldsymbol{HR} = \boldsymbol{U} \qquad 9\text{-}18$$

R is an orthogonal matrix with columns of $\boldsymbol{r}_i$ -eigenvectors of the Hessian matrix. ***U*** is the diagonal matrix with the corresponding eigenvalue elements $U\left(\boldsymbol{r}_i\right) - U\left(\boldsymbol{r}_i^0\right)$. The 3*N* ***r***-eigenvectors, which are the normal modes, are **linearly independent** and thus form a **basis** for the R^{3N} space as mentioned in the previous sections (Edwards and Penney 1987), pp. 164).

[a] Vectors are linearly independent if they are not linear combinations of each other.

[b] An eigenvalue matrix is a diagonal matrix with eigenvalues along the diagonal. An eigenvector matrix is a square matrix with columns of linearly independent eigenvectors. An eigenvector in the eigenvector matrix is associated with the eigenvalue in the eigenvalue matrix with the same column number. Edwards, C. H. and D. E. Penney (1987). Elementary Linear Algebra., p.270-274

[c] "The n x n matrix A is diagonalizable if and only if it has n linearly independent eigenvectors" Ibid., p. 273

[d] The square matrix P is per definition orthogonal if $P^{-1} = P^{T}$. Ibid., p.292

[e] "A square matrix is orthogonally diagonalizable if and only if it is symmetric." Ibid., p.296

9.2.3.5 Gaussian networks

Sections 9.2.3.2-9.2.3.4 outlined how to calculate the elements of the Hessian matrix and how to diagonalize the Hessian matrix. Generally, one would assume atoms in a protein are connected only if covalently bonded to each other. However, this is not the case for Gaussian networks, which are the basis of all normal mode calculations in this thesis. In this section Gaussian networks are introduced.

A Gaussian chain is a freely jointed chain. The joints and chains occupy zero volume and are thus not excluding each other in space. Freely jointed refers to the fact that the distance and orientation between joints is not fixed. A Gaussian phantom network is thus nothing but a network of independently flexible chain segments. Gaussian refers to the fact that the number of possible chain segment conformations follows a Gaussian function (i.e. the Gaussian segment-configuration function)(James 1947). In the phantom network the flexibility of the Gaussian chains is only constrained by the connectivity of the joints.(Flory 1976)

When applying the Gaussian phantom network model to proteins, only spatial connectivity of atoms is taken into consideration, as compared to constraints from for example electrostatic and steric interactions which are not. The three dimensional Gaussian network for a protein is illustrated in Figure 59. In the Gaussian network model a phantom spring between atoms is only present, if the distance between the atoms is less than a **cutoff distance** value. Otherwise the second derivatives in the Hessian matrix corresponding to these two atoms are set equal to zero. The equation below shows which factor to multiply the second derivatives with when using a normal "hard" cutoff. In section 9.2.3.10 the sigmoid cutoff will be introduced.

$$H_{ij} = \begin{cases} -1 & s_{ij} < s_c \\ 0 & s_{ij} > s_c \\ -\sum_{i \neq j}^{N} H_{ij} & i = j \end{cases} \tag{9-19}$$

The cutoff distance should at least be set high enough to avoid multiple zero value eigenvalues. The extra zero eigenvalues will correspond to movements of autonomous networks relative to one another, which are not associated with a cost in energy.

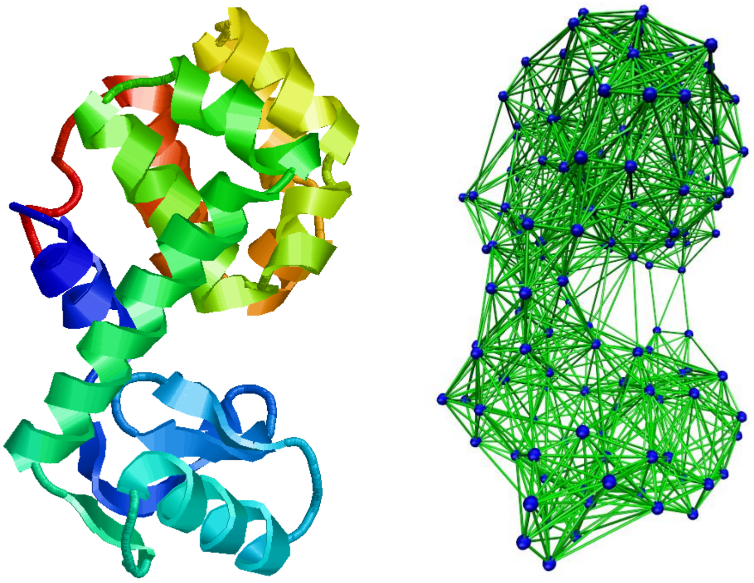


Figure 59 – Illustration of a Gaussian network in T4L (PDB ID 2lzm). On the left the protein secondary structure is represented by cartoon helices and strands. On the right the Gaussian network of the protein is displayed using a cutoff distance of 8Å. Only connections between C_α atoms are shown. Drawn with VMD.(Humphrey, Dalke et al. 1996)

9.2.3.6 Overlap

Eigenvectors are the result of diagonalizing the Hessian matrix and a conformational change is described by a vector. It is therefore desirable to be able to compare the similarity of two vectors. In R^3 space two vectors overlap if they are parallel and point in the same direction. The cosine of the angle θ between the vectors r_1 and r_2, which here is called the overlap, is equal to the dot product of the two vectors divided by their length

$$\cos\theta = \frac{r_1 \cdot r_2}{|r_1||r_2|} \qquad 9\text{-}20$$

In R^3 space with 3 dimensions the overlap is geometrically interpretable. To maintain a geometric interpretation, the overlap between two eigenvectors in R^{3N} space could be calculated using the average overlap of each of the N residues. But the above formula is used even in R^{3N} space(Marques and Sanejouand 1995). The geometric interpretation of the overlap is infrequently retained by calculating an average of the overlap between vectors of individual atoms.(Krebs, Alexandrov et al. 2002)

9.2.3.7 Physical interpretation of the eigenvalues and the eigenvectors of the Hessian matrix

The physical interpretation of the value of the eigenvalues is related to the curvature of the potential energy surface. Moving along an eigenvector that has a large eigenvalue $U(\boldsymbol{r}_i) - U(\boldsymbol{r}_i^0)$ is associated with a large cost in energy. When adding thermal energy to the system consisting of the protein, then the lowest frequency modes are most likely to be populated. The distribution of movements follows a Boltzmann distribution. Movements in directions of low potential energy are more likely to occur. Due to Taylor expansion around a minimum $U\left(\boldsymbol{r}_i^0\right)$ of the potential energy function, the quadratic form and the Hessian matrix are positive (semi)definite[a]. This means that all eigenvalues are either zero or positive. Furthermore none of the eigenvalues from diagonalization of the Hessian matrix are complex.[b]

The potential energy is only dependent on the coordinates. Translating or rotating the structure in space does not change the relative position of the coordinates to each other and thus the potential energy is not changed. Therefore the six eigenvalues corresponding to the eigenvectors of translation and rotation along the axes in R^3 space are expected to be zero.

A linear combination of the linearly independent eigenvectors can describe any initial movement upon addition of thermal energy. Nature travels along a path of low energy because conformational states of different energy are populated according to the Boltzmann distribution. Protein dynamics is not expected to be different. It can be shown that the conformational change between an open (pdb 150L) and closed (pdb 2LZM) conformation of T4 lysozyme is better described by the eigenvectors associated with the low eigenvalues as can be seen from Figure 60. The normal mode calculations were based on the open conformations, since it was previously shown that overlaps are higher when the transition is from open to closed conformation(Tama and Sanejouand 2001), which was also the case for the two conformations of T4 lysozyme selected here.

And indeed, Gerstein and his colleagues(Krebs, Alexandrov et al. 2002) demonstrated from a set of proteins available in two conformations that the eigenvectors associated with the lower nonzero eigenvalues are often the best at describing the motion between the two super imposed conformations. In other words, the maximum overlap with the conformational difference vectors is most often observed for the eigenvectors associated with the lowest nonzero eigenvalues as can be seen from Figure 61.

The large scale results of Gerstein and Krebs validate the use of the eigenvectors associated with low nonzero eigenvalues for the description of directions of conformational changes.

[a] "Let $q(\mathbf{x}) = \boldsymbol{x}^T A\boldsymbol{x}$ be a quadratic form with symmetric n x n matrix A. Then q is positive definite if the eigenvalues of A are all positive." (Ibid., p.369)

[b] "The characteristic equation of a symmetric matrix has only real solutions." Ref. Ibid.

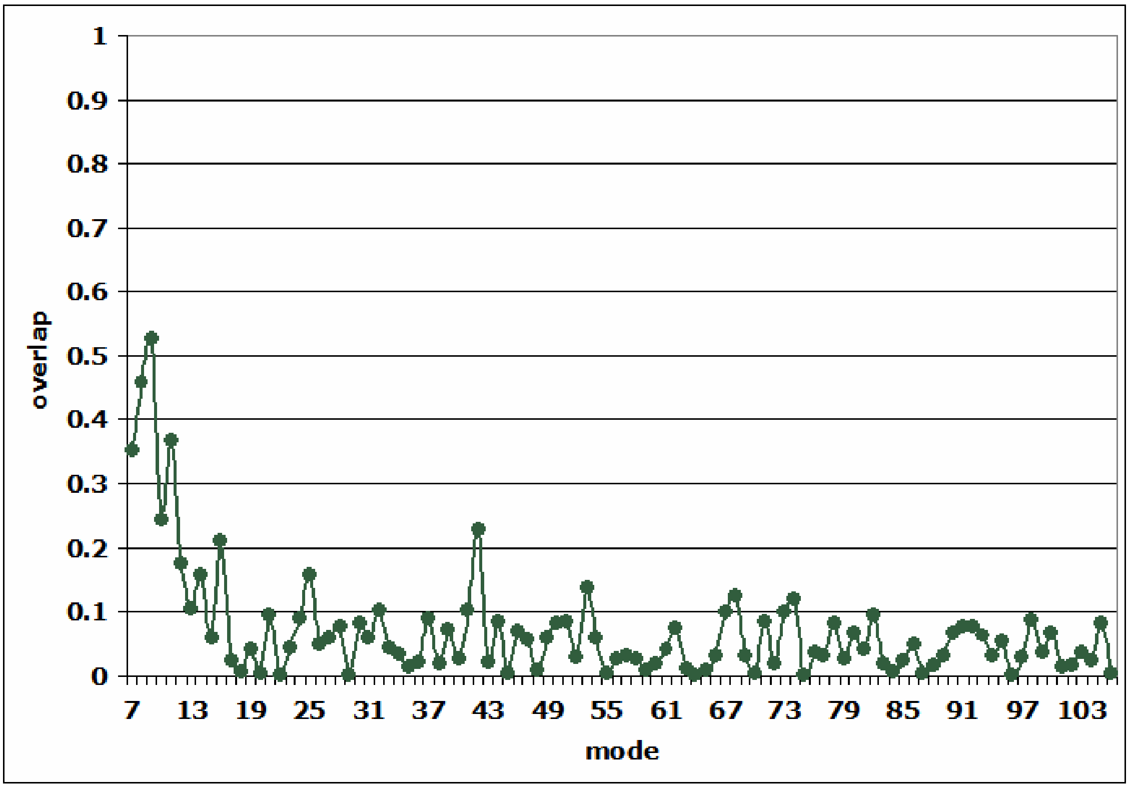


Figure 60 – Plot of the 100 first normal modes of T4 lysozyme in a closed state (2LZM) versus the overlap of the normal modes with the vector of the conformational change to the open state (150L).

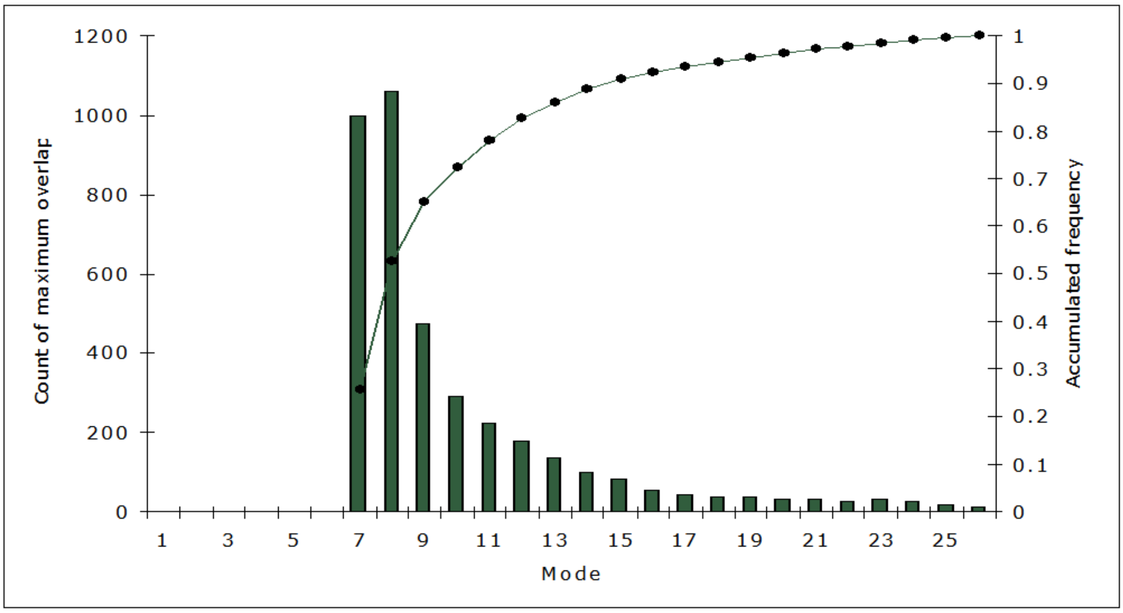


Figure 61 – Distribution of the mode of maximum overlap for a selection of more than 3000 proteins available in two conformations. The overlaps are calculated between the vector of conformational change and the eigenvectors of individual normal modes. The mode on the abscissa is plotted against the count of maximum overlap on the left ordinate and the cumulated frequency on the right ordinate. The figure is modified from (Krebs, Alexandrov et al. 2002).

9.2.3.8 Simplified force field

The energy of a protein can be calculated using empirical force potentials involving terms for covalent interactions (e.g. bond stretching, angle bending, bond rotation) and noncovalent interactions (e.g. electrostatic interactions of charges and dipoles and steric interactions) as known from molecular dynamics simulations.(Tirion 1996) From 1979 to 1996 classical normal mode analysis was solely based on full force field calculations. This involves time consuming energy minimization of a structure and calculation of full force field energies and their second derivatives for use in the Hessian matrix. Then, in 1996, Tirion(Tirion 1996) reproduced the relative residue fluctuations of full force potentials using a simplified force field and connectivity network. In Tirion's elastic network model, the energies are dependent only on a single uniform force constant and the distance between atoms. Since then, even more advanced connectivities involving covalent backbone interactions, disulfide bonds and electrostatic interactions have been introduced.(Jeong, Jang et al. 2006) However, by randomization of matrix elements, it has demonstrated that eigenvectors are determined by the **shape** of the Hessian matrix rather than the values of the **elements** of the Hessian matrix(Lu and Ma 2005); i.e. the coordinates of a protein structure are more important than e.g. charges, dipoles and disulfide bonds. Previously it was also established that global conformational changes rely more on overall structure rather than structural details.(Kitao and Go 1999) It is therefore well documented that coarse grained methods are appropriate for describing large scale conformational changes. There is no immediate need for advanced force fields. Therefore upon building the Hessian matrix a uniform force field is used and all connections of identical length are considered equally strong. The structure is not energy minimized prior to the calculations, since the method is only based on the positions of the C_α atoms as explained in the next section.

In this thesis all calculations are performed in vacuum. Normal mode calculations in general never include the solvent. My algorithm is no different. It is important to note here that not factoring the solvent into the calculations may impact the results, if, as proposed, solvent fluctuations dominate protein dynamics.(Fenimore, Frauenfelder et al. 2002)

9.2.3.9 Coarse graining

The number of row operations required for calculating the eigenvectors of an NxN matrix increases by the cube of N and the number of matrix elements increases by the square of N. Thus, memory and processor requirements for diagonalization of the NxN matrix increase by N^2 and N^3 respectively. Even though memory might not limit the calculation, it is still of interest to reduce the dimensions of the Hessian matrix to speed up calculations. This can be done by using only C_α atoms(Hinsen 1998; Hinsen 1999), residue mass centers(Hinsen 1998), or even blocks of residues(Durand P 1994) as the interconnected nodes instead of individual backbone and sidechain atoms. For all results in this thesis, the method of C_α coarse graining has been employed. Coarse graining reduces the number of degrees of freedom – those of the side chains - but the eigenvectors associated with the low nonzero eigenvalues still succeed in describing X-ray B-factors(Bahar, Atilgan et al. 1997; Haliloglu, Bahar et al. 1997; Doruker P 2000) and the motion between open and closed conformations(Tama and Sanejouand 2001).

9.2.3.10 Sigmoid cut-off and Hessian matrix calculation

Sequence identical structures obtained at different temperature, ion concentration or pressure or by different methods might have slightly different structures. To avoid getting different results for nearly identical structures a sigmoid cutoff distance is used instead of a sharp cutoff distance as recommended by Lynn Ten Eyck. Previously, Hinsen has used an exponential cutoff.(Hinsen 1998) Here I use a sigmoid cutoff. The equation used is derived from that of a sigmoid function.

$$y = \frac{1}{1 + e^{x-10}} \tag{9-21}$$

x is the distance between C_α atoms and *y* is the factor with which the uniform force constant is multiplied. A plot of the function is shown in Figure 62.

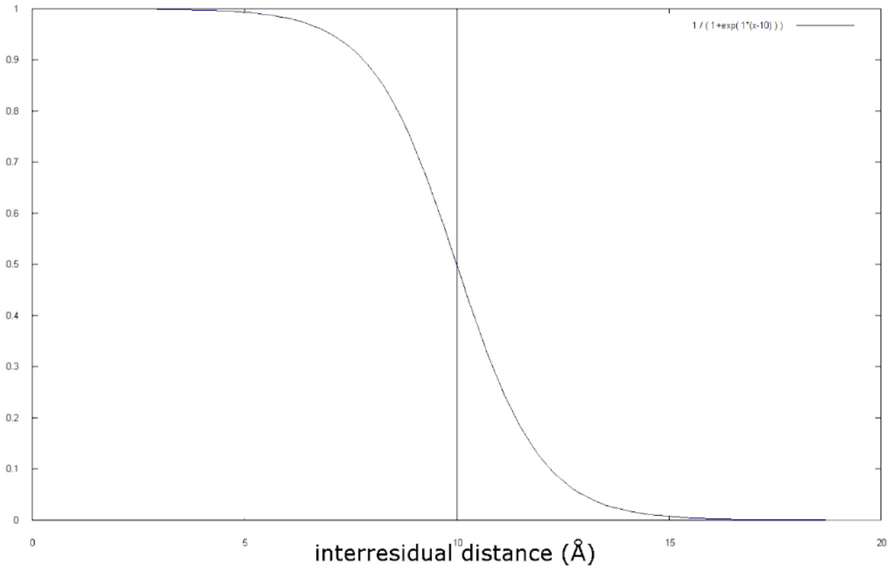


Figure 62 – A plot of a function derived from the sigmoid function. On the abscissa is the distance between junctions and on the ordinate is the number with which the force constant is set to be proportional to. From the plot a smooth transition in the area around 10Å is seen.

The most essential part of normal mode analysis is the calculation of the elements of the Hessian matrix. Each element of the Hessian matrix is calculated by using equation 9-8. Diagonal elements are calculated by using equations 9-9 and 9-10. Each element is then multiplied by the sigmoid factor, *y*, which is calculated from equation 9-21. Once the Hessian matrix has been built it can be diagonalized and the eigenvectors and eigenvalues retrieved.

9.2.3.11 Importance of the input structure

To determine if the conformational variation between structures is of any significance to the calculated normal modes, I have calculated the normal modes of 63,810 sequence identical monomeric protein pairs from the same space group. I have then calculated the overlap between normal mode 7 of each of the two proteins. The overlap I have then plotted against the C_α RMSD between the two structures (Figure 63). The calculated normal modes are almost identical (overlap > 0.98) for most of the 63,810 structure pairs, and the overlap probably only drops as a function of RMSD as a consequence of different origins of the calculated eigenvectors. In most cases (>99%) the specific conformation of the input structure is not of importance to the calculated normal modes. The method is robust.

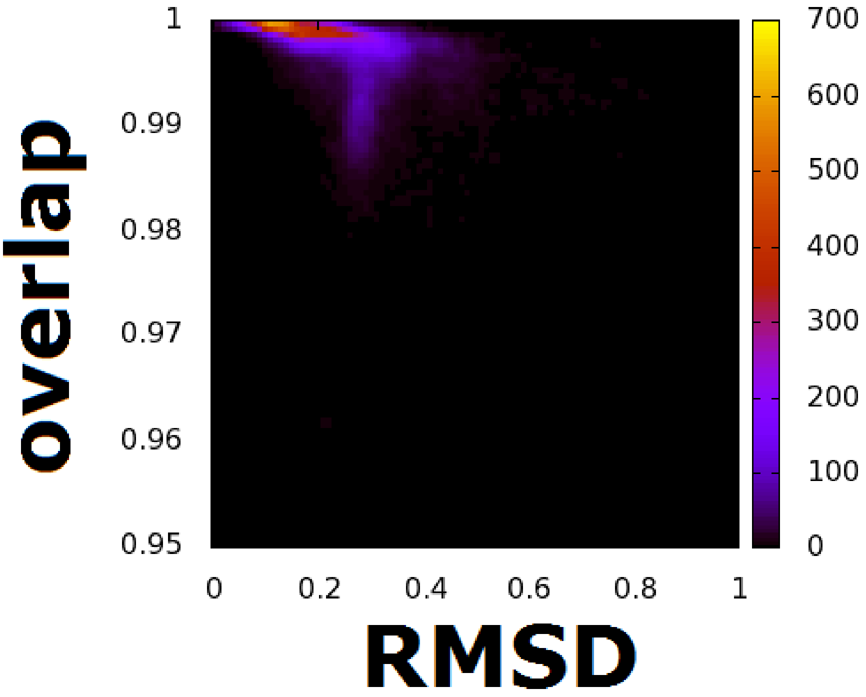


Figure 63 – The overlap between normal mode 7 of two sequence identical structures as a function of the C_α RMSD between them.

9.2.4 Cavity finding by addition of mass centers and recalculation of normal modes

After having explained the theory of normal mode analysis in the previous section I now explain how my NMA algorithm for ligand binding site identification works. The basis of an NMA calculation is a network of springs between atoms as explained previously in section 9.2.3.5. The calculation can be coarse grained by only selecting C_α atoms as shown for T4L in blue in Figure 64a. From Figure 64b one can relatively clearly see, that the easiest way of distorting the network of springs is to move the upper and lower cluster of points (the two helix-connected domains of the protein) towards or away from each other. An NMA calculation arrives at the same conclusion (Figure 64c), but identifies the movement by performing a set of mathematical operations on a matrix describing the spring connections between points.

In Figure 64c the vectors of the normal mode associated with the smallest change in energy (normal mode 7; the motion that is the easiest to perform) is mapped onto the network of points (T4L C_α atoms). This motion correlates with the differences observed among the many different conformations of T4L recorded by X-ray crystallography as I explained in section 9.2.3.7. The motions are compared by measuring the overlap between the calculated vectors and the vectors describing the differences between two X-ray structures. In the case of T4 lysozyme the second lowest normal mode (the second easiest motion to perform) has the highest correlation with the conformational change observed in X-ray structures. Whereas the first normal mode merely describes a hinge motion, the second normal mode also involves a shear motion around the long helix running from Lys60 to Arg80. The second normal mode therefore has a higher correlation with the conformational change, which is not a perfect hinge motion. By studying a large set of proteins, it has been shown that the lowest normal modes generally correlate with experimentally observed conformational changes.(Gerstein and Krebs 1998)

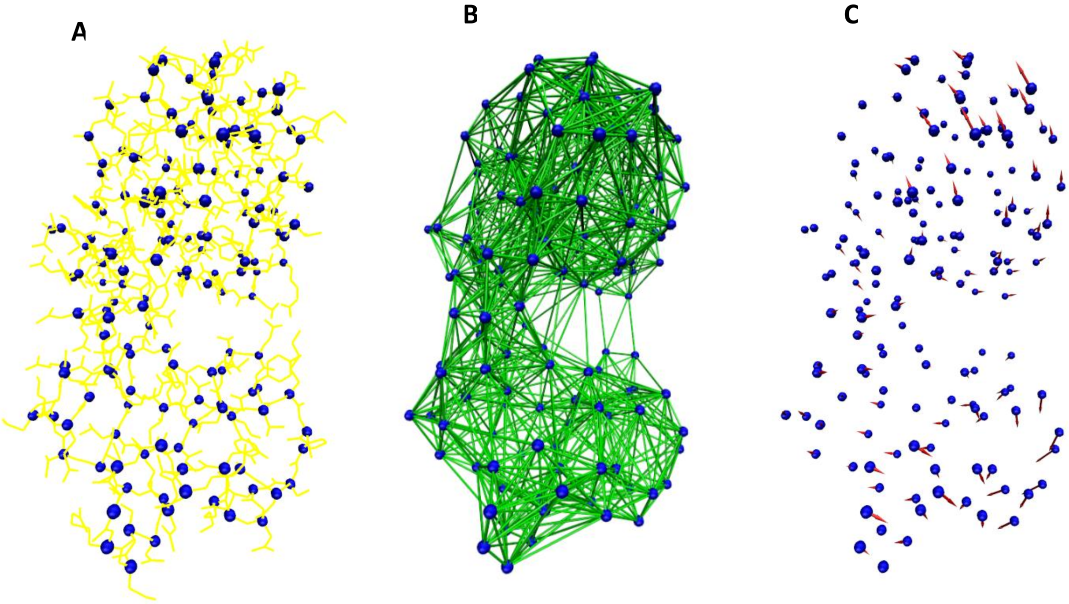


Figure 64 – Different representations of T4 lysozyme (2lzm). C_α atoms are shown as blue spheres. On the left a yellow stick model of the enzyme is shown. Shown in the middle are green springs between C_α atoms within 10Å of each other. Shown on the right for one mode are red vectors originating from C_α atoms and pointing in the direction of movement calculated with NMA. The largest vectors correspond to the largest mobility.

I decided to apply normal mode analysis to the problem of identifying ligand binding sites. I simulate the presence of a ligand by adding an extra interaction point at the surface of the protein inspired by a previous similar method applied to the problem of identifying sites of nicotinic acetylcholine receptor, which can alter the gating mechanism of the protein.(Taly, Corringer et al. 2006) The presence of a ligand will often perturb the structure of a protein significantly(Bakan and Bahar 2009). Blocking the active site would prevent the protein from sampling the ligand bound form. By simulating a ligand at the surface and determining what positions cause the largest reduction in overlap with the eigenvector calculated from the ligand free form I can therefore identify ligand binding sites.

I do not expect to perform well upon identification of internal cavities, as the atom density in this case is very high. Adding a probe atom to an already highly dense area will not alter the calculated eigenvalues and eigenvectors, as the values of the Hessian matrix will change insignificantly, as they are already non-zero prior to perturbation. It has been shown that it is the shape of the Hessian matrix rather than the values that constitute it, which decides the outcome of the diagonalization of the matrix.(Lu and Ma 2005)

My NMA method predicts ligand binding sites by a multi step process. The first step is the construction of a grid of points surrounding the protein. A grid spacing of 2Å is used, which increases the calculation speed, while maintaining an adequate precision. Points further away than 6Å from and closer than 3Å to any C_α atom of the protein are not considered. At first the normal modes of the protein are calculated with no additional mass center present and then recalculated upon placing a mass center at each of the grid points. Next a mass center is placed sequentially at each grid point and the difference in the dominant mode (mode 7) is calculated for each position of the mass center. The ligand binding site is identified as the grid point(s) where the addition of a mass center gives the largest change in the directionality of the dominant normal mode.

It has been shown that perturbation of the Hessian matrix increases the deviation between higher normal modes more than that between lower normal modes, when normal modes calculated from the original and perturbed Hessian matrix are calculated.(Lu and Ma 2005) The higher eigenvalues are also less spaced and “high” normal modes therefore swap more easily. Therefore it is justified to compare perturbed normal modes, if the comparison is limited to normal modes with a low index, like I do here.

9.2.5 Selection of a structure test set

In addition to the four model proteins, a set of 99 sequence identical ligand free/bound structure pairs from the PDB formed the benchmarking set for my NMA method. This set was created by first selecting all monomeric X-ray structures with one bound ligand in the PDB. I chose to only use monomeric proteins, so the problem does not expand to one of finding ligand binding sites at multimer interfaces. I accepted structures with missing residues at the terminals, but I excluded structures which had residues missing in the middle of a chain. I accepted zero occupancy residues. I only accept ligands with 10 or more atoms and I do not consider ligands such as peptides, nucleotides, ions, prosthetic groups, co-enzymes and various other solutes (see appendix). After this exclusion the redundant list contained 2784 ligand bound structures of which many are sequence identical. Next BlastClust(Altschul, Gish et al. 1990) was used to find ligand free structures that are 100% identical in sequence to one of the ligand bound structures. This reduced the number of ligand free and bound structures to 2234 and 1825 respectively, distributed across 581 clusters of sequence identical proteins. Next a non-redundant dataset in terms of protein sequence was created. I expected my NMA method to perform better in those cases, where the motion between apo and holo structure was described by normal mode 7. Therefore I chose a ligand free and bound structure from each cluster based on which free/bound motion was most similar to normal mode 7. For a further reduction of the size of the dataset, structures not present in a database of biologically relevant binding sites (version 9.4 of LigASite)(Dessailly, Lensink et al. 2008) were excluded from the final dataset. This procedure yielded a dataset of 99 sequence identical free/bound structure pairs. My NMA algorithm for ligand binding site identification described in the previous section (9.2.4) is then applied to this dataset. The results are presented in section 9.3.3.

9.2.6 Normalized residue B-factors and Jensen-Shannon divergence

In section 9.3.4 I show that conserved residues like those in binding sites rarely are very flexible as measured by their B-factors. In section 9.3.4 I make use of the "normalized residue B-factors" and sequence conservation scores. Here I present the calculation of each of the parameters.

A "residue B-factor", B_{res}, I simply define as the average B-factor of the backbone atoms of that residue. The averaged B-factors are then normalized in order to exclude differences in B-factors between structures due to different resolutions(Rupp 2009) and different hydration levels (see appendix).

$$B_{\text{res,normalized}} = \frac{max(\boldsymbol{B}_{\textbf{res}}) - B_{\text{res}}}{max(\boldsymbol{B}_{\textbf{res}}) - min(\boldsymbol{B}_{\textbf{res}})}$$

The measure of sequence conservation that I use is the Jensen-Shannon divergence.(Capra and Singh 2007) The Jensen-Shannon divergence (JSD) is a measure of the dissimilarity between multiple sequences. It ranges from 0 to 1. The higher the score, the more conserved the residues is. The JSD is derived from the Kullback–Leibler divergence (KLD), which measures the difference between two probability distributions *P* and *Q* for each of the 20 amino acids, *aa*.

$$D_{\text{KL}}\left(P \,||\, Q\right) = \sum_{aa}^{20aa} P\left(aa\right) \log \frac{P\left(aa\right)}{Q\left(aa\right)} \qquad (9\text{-}22)$$

P is the frequency of occurrence of an amino acid in a column in a multiple sequence alignment. *Q* is a background distribution of amino acids. One reason I prefer the JSD over the

KLD is that it is bound by the values 0 and 1(Lin 1991) and thus normalized, whereas there is no pre-defined upper limit for the KLD. The relation between the JSD and the KLD is

$$D_{JS}\left(P\,||\,Q\right)=\frac{1}{2}D_{\mathrm{KL}}\left(P\,||\,\frac{1}{2}\left(P+Q\right)\right)+\frac{1}{2}D_{\mathrm{KL}}\left(Q\,||\,\frac{1}{2}\left(P+Q\right)\right) \qquad (9\text{-}23)$$

The JSDs were downloaded from the internet (http://compbio.cs.princeton.edu/concavity/pqs/jsd). The JSDs were calculated on the HSSP dataset of multiple sequence alignments(Dodge, Schneider et al. 1998) using the BLOSUM62 matrix as the background distribution reference *Q*(Henikoff and Henikoff 1992) and.

Only monomeric X-ray structures are analyzed. Structures with identical B-factors for each atom are excluded from the analysis. Modified residues are skipped, but structures with modified residues are not excluded from the analysis. In total 3,340,239 residue B-factors and JSDs from 12,463 proteins are plotted.

9.3 Results

9.3.1 Evaluation of the NMA algorithm

Before my NMA algorithm can be applied to the problem of ligand binding site prediction, it must be thoroughly tested and its performance compared to other NMA algorithms. Before I present the results of the ligand binding site prediction, I present the performance of my NMA algorithm.

In the current algorithm, which makes use of a uniform force constant, the distance cutoff is the only parameter, which can be changed. It is therefore critical to show that similar results are obtained within a range of cutoff distances. This is done in section 9.3.1.3 and in section 9.3.1.3 it is shown that the use of a cutoff distance of 10Å is an appropriate choice for modeling the conformational change between two conformations of T4 lysozyme (2LZM and 150L:D). The amplitudes calculated with my NMA algorithm will be shown to correlate with X-ray B-factors (section 9.3.1.1) and the vectors between NMR ensembles (section 9.3.1.2). The former is a standard method to validate normal mode analysis(Bahar, Atilgan et al. 1997; Haliloglu and Bahar 1999), whereas the latter is a novel method of validating the results of normal mode analysis.

9.3.1.1 Correlation with isotropic X-ray B-factors

The algorithm developed has proven to be equally good compared to other algorithms(Bahar, Atilgan et al. 1997; Haliloglu and Bahar 1999) at predicting X-ray temperature factors. The correlation between B-factors of C$_\alpha$ atoms and normalized amplitudes of selected modes are shown in Figure 65. If it is assumed that the B-factors and the calculated amplitudes follow a normal distribution, then *t* test statistics can be calculated to determine, whether the data are correlated or not.(Zar 1998) The results of the statistical *t* test for the correlation between modes 7-3*N* and the temperature factors are shown in Table 15. Only for 2OMF do the calculated amplitudes not correlate with the B-factors. The lack of correlation might be explained by the fact that 2OMF in the crystal is present as a trimer, whereas the calculations are based on the monomer.

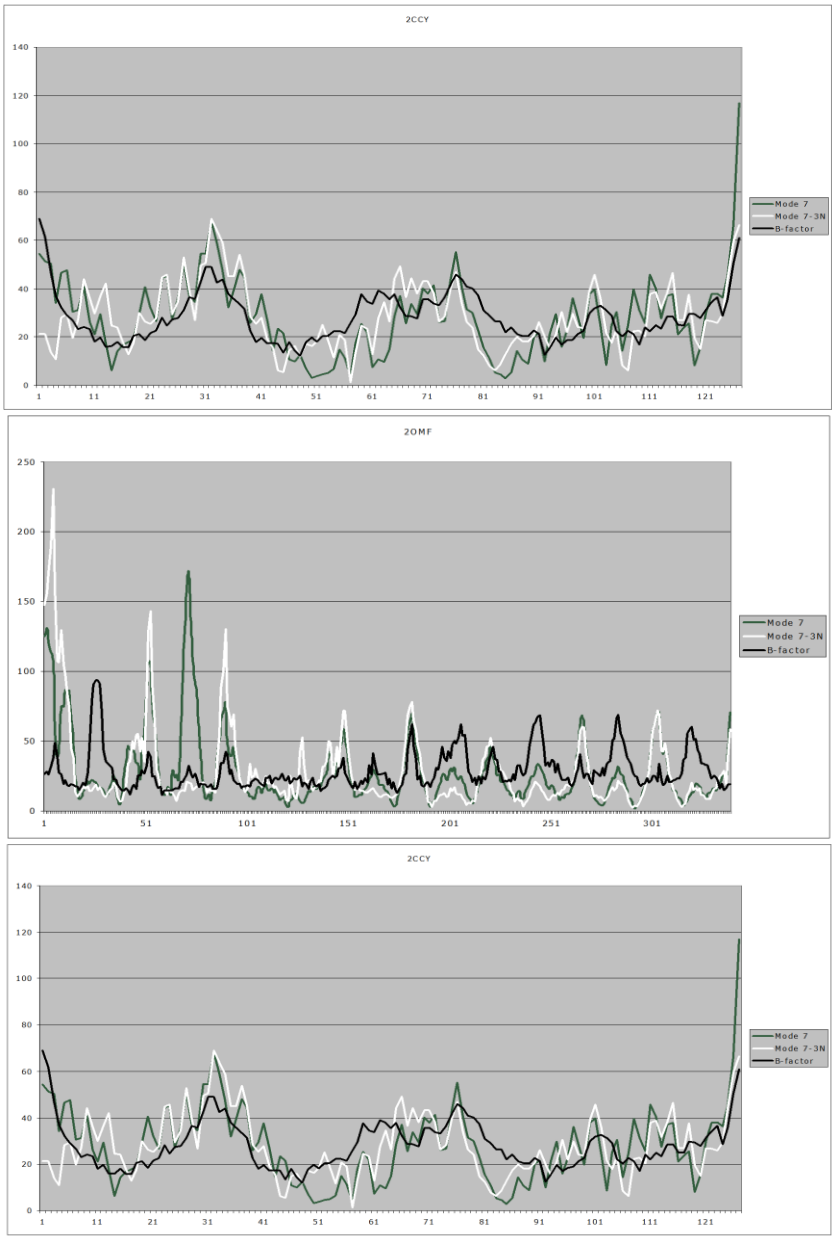


Figure 65 – Correlation between the B-factors (black) of the C_{α} atoms of cytochrome C (2CCY), Ompf Porin (2OMF) and T4 lysozyme (3LZM) and the amplitudes of mode 7 (Chargaff, Lipshitz et al.) and the linear combination of modes 7-3N (white).

PDB	correlation (r)	degrees of freedom (n-2)	t statistic	probability (p)
2CCY	0.48	125	6.13	< 0.001
2OMF	0.04	338	0.77	0.45
3LZM	0.47	162	8.31	< 0.001

Table 15 – Correlations between X-ray temperature factors and the calculated amplitudes for modes 7-3N. The probability of correlation is calculated by the statistical *t* test.

9.3.1.2 Correlation with NMR fluctuations

Previously NMR S^2 relaxation parameters have been compared to the fluctuations calculated by normal mode analysis.(Haliloglu and Bahar 1999) The equilibrium fluctuations between ensembles of NMR structures have never been compared to the fluctuations of normal mode analysis though. In Figure 66 the correlation between the two normalized amplitudes are shown. The NMR amplitudes, *A*, for individual residues, *i*, are calculated by averaging the amplitude between all combinations, *j,k*, of all the NMR ensembles of a structure deposited in the protein data bank. The equation in use is shown below.

$$A_i = \sum_j^n \sum_k^n \sqrt{\left(x_j - x_k\right)^2 + \left(y_j - y_k\right)^2 + \left(z_j - z_k\right)^2} \qquad 9\text{-}24$$

Again *t* test statistics were calculated to check if the correlation was significant or not. The normal mode amplitudes were calculated from a linear combination of the first 20 modes. The results are shown in Table 16. It cannot be rejected that there is no correlation between the amplitudes for 2D21. From Figure 66 it can be seen that the correlation is poor for especially residues 1-110 and 261-330. Both residue ranges map to the immobile region of the protein. If the correlation is calculated for residues in the ranges 111-260 and 331-374 then $r = 0.58$ and p < 0.001. And for residues 111-260 the correlation is as high as 0.73.

I also check my algorithm to see if it is the low energy normal modes that in general are the best at describing NMR fluctuations. I plot the correlation between the NMA RMSF and the NMR RMSF against the index of the normal mode. What is really interesting is that the correlation does not level out at zero for higher modes, as one would expect if the high energy normal modes were random motions. Instead the correlation is a linear function of the normal mode index and there is anti-correlation for the high energy normal modes rather than no correlation at all (appendix). This is another strong indication to me, that focusing only on low energy normal modes - when describing conformational changes and finding ligand binding sites - is the right approach.

PDB	correlation (*r*)	degrees of freedom (n-2)	*t* statistic	probability (*p*)
1BNR	0.87	108	25.07	< 0.001
1E8L	0.62	127	11.35	< 0.001
2D21	0.06	162	1.23	0.22

Table 16 – Correlations between NMR amplitudes and the normal mode amplitudes for modes 7-26. The probability of correlation is calculated by the statistical *t* test.

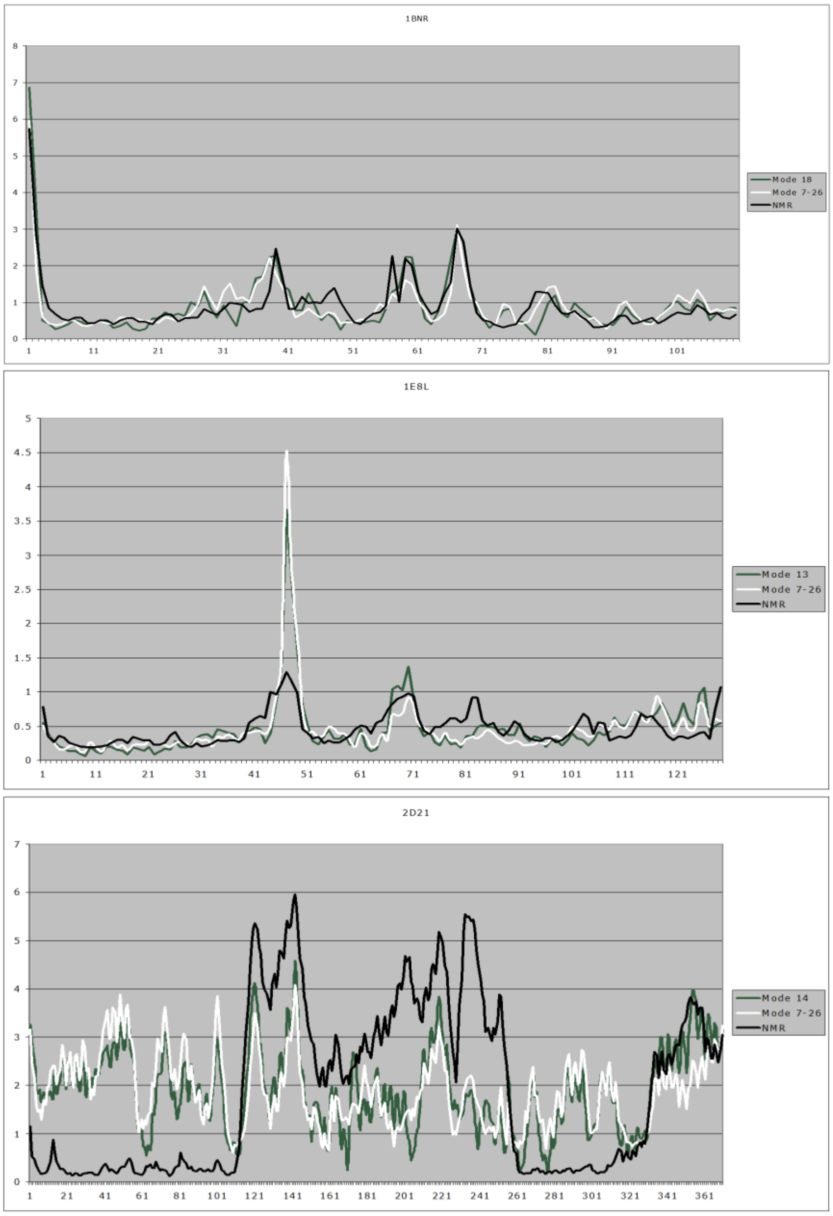


Figure 66 – Correlation between the NMR amplitudes (black) of the C_α atoms of Barnase (1BNR), Hen Egg White Lysozyme (1E8L) and Maltodextrin-Binding Protein (2D21) and the amplitudes of a single mode (Chargaff, Lipshitz et al.) and the linear combination of modes 7-26 (white).

9.3.1.3 Variability of the cutoff distance

When using a uniform force constant and a coarse grained C_α Gaussian network model, only the distance cutoff affects the results of the normal mode analysis. It is important that the differences in the results are not significant and influence the conclusions drawn from the normal mode analysis upon changing the distance cutoff. To investigate this one can change the distance cutoff in small steps and investigate the changes in for example overlap with a conformational change vector.

This is illustrated in Figure 67 for the conformational change between T4 lysozyme in its open (150L, chain D) and closed conformation (2LZM). The coordinates used for calculation of the Hessian matrix are those of the closed conformation. The cutoff distance was changed in increments of 0.1Å in the range from 0.0Å to 50.0Å, which is just slightly less than the diameter of T4 lysozyme. A cutoff distance of 50.0Å will thus cause most atoms in T4 lysozyme to be connected in the network. Starting at the left of the figure it is seen that overlaps are constant until a threshold of approximately 5Å and then start to vary significantly. The reason is that the minimum distance between C_α atoms is approximately 5Å-5.5Å, which is a value that can be calculated by averaging the distance to the nearest C_α for all C_α atoms in a protein. As has been previously mentioned the Hessian matrix will yield multiple zero eigenvalues upon diagonalization if the network is disconnected and split into autonomous networks. The maximum overlap and the average overlap of modes 7-12 is seen to be roughly constant in the interval from 10Å to 25Å. This supports the choice of 10Å as a cutoff distance in the case of T4 lysozyme. Since an analysis of variability of the cutoff distance has only been done for T4 lysozyme, it is assumed that the use of this cutoff distance can be extended to other systems of different size and structure. In the literature a cutoff distance in the range 10Å-15Å is the most common. This concludes my evaluation of my NMA algorithm, and the results on ligand binding site identification follow.

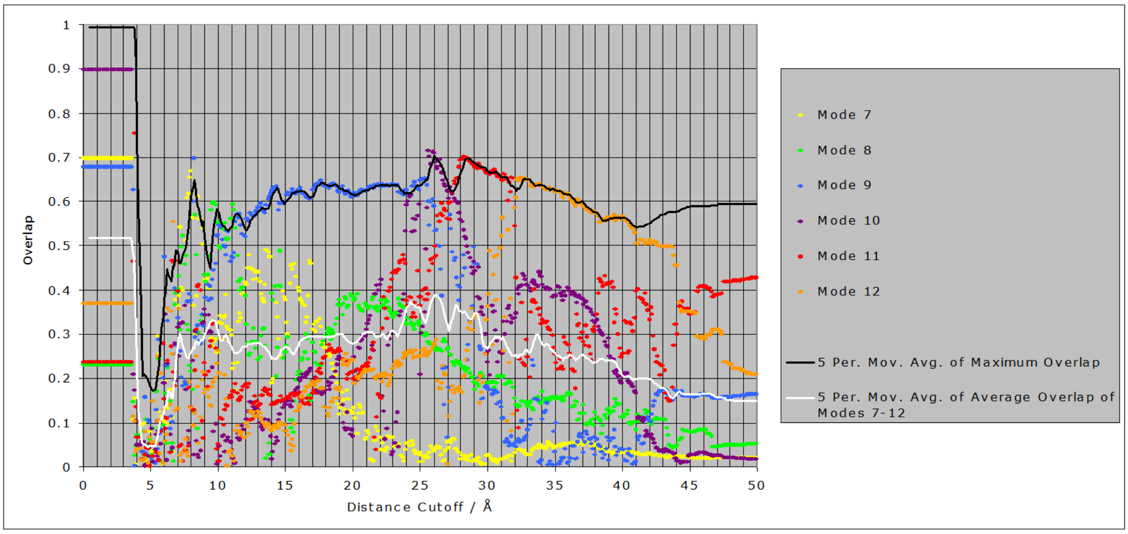
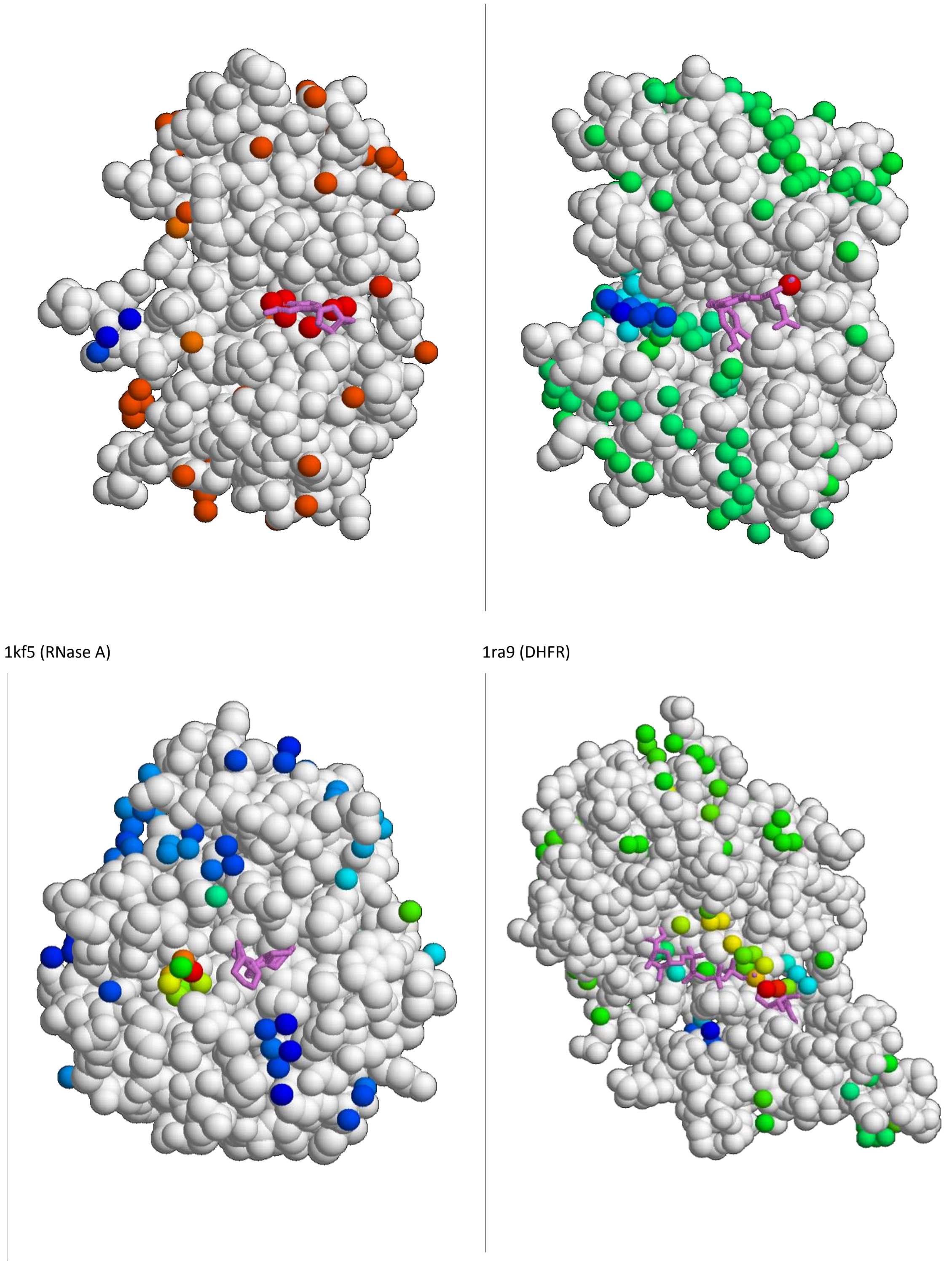


Figure 67 – The overlap of normal modes with the 2lzmA-150lD-vector as a function of cutoff distance (Å) for individual modes. Shown in black is the 5 period moving average of maximum overlap among all modes. Shown in white is the 5 period moving average of the average overlap among mode 7-12.

9.3.2 Identification of the ligand binding site in my four model proteins

I now turn my attention to the identification of ligand binding sites in the four model proteins that bind their ligands by conformational selection. As already explained normal modes should describe conformational changes at a pre-exisiting equilibrium with no energy barriers. The addition of nodes to the elastic network should have the largest effect on the calculated normal modes if those nodes are placed at the ligand binding site; thus enabling the identification of the ligand binding site. For all four proteins the identification of the ligand binding site is successful (Figure 68).



1kf5 (RNase A)

1ra9 (DHFR)

1w8v (CypA)

2rh5 (AdK)

Figure 68 – Identification of ligand binding sites in RNase A (top left), DHFR (top right), CypA (bottom left) and AdK (bottom right). The ligand is shown in violet. The probe positions causing the largest decrease in overlap are shown in red and the smallest decreases in overlap are shown in blue. For all four proteins binding their ligand by conformational selection, the ligand binding site is correctly predicted; i.e. the probe position causing the largest decrease in overlap (shown in red) is within 3Å of the nearest ligand atom (shown in violet).

Whether conformational sampling is likely to be of importance to the free/bound motion or not can be determined by determining the level of contribution of each of the calculated normal modes to the free/bound motion. If conformational sampling is likely to funnel the free/bound conformational change, then a large contribution from the lowest normal modes towards the free/bound motion would be expected, and otherwise all of the normal modes would be expected to contribute equally to a description of the free/bound motion. For the four model proteins the low energy normal modes contribute the most to the free/bound motion (Figure 69). On the contrary in beta-lactoglobulin there are no dominant modes contributing to the free/bound motion (Figure 70); either that or the free/bound motion is not a functionally relevant motion and/or the free/bound motion is not well described by the selected free and bound structures. Interestingly conformational changes in beta-lactoglobulin is known to happen due to outside effects (pH changes)(Sakurai and Goto 2007), whereas the conformational changes in the four model proteins is an intrinsic/spontaneous event, which is believed to be of importance to the ligand binding properties of the proteins. In CypA and DHFR the free/bound motion is less well defined by a few normal modes than it is in AdK and RNase A. This might be because of the experimental errors of the placement of atoms, which "hides" the real motion between the free and bound structure. With C_α RMSDs of less than 0.2Å this is not unlikely. The C_α RMSD between the two AdK structures is more than 5Å, and it is therefore expected that normal mode 7 will contribute the most to a description of the motion, if the conformational change between the free and bound structure is spontaneous at a pre-existing equilibrium and involves no crossing of any significant energy barriers (Figure 69).

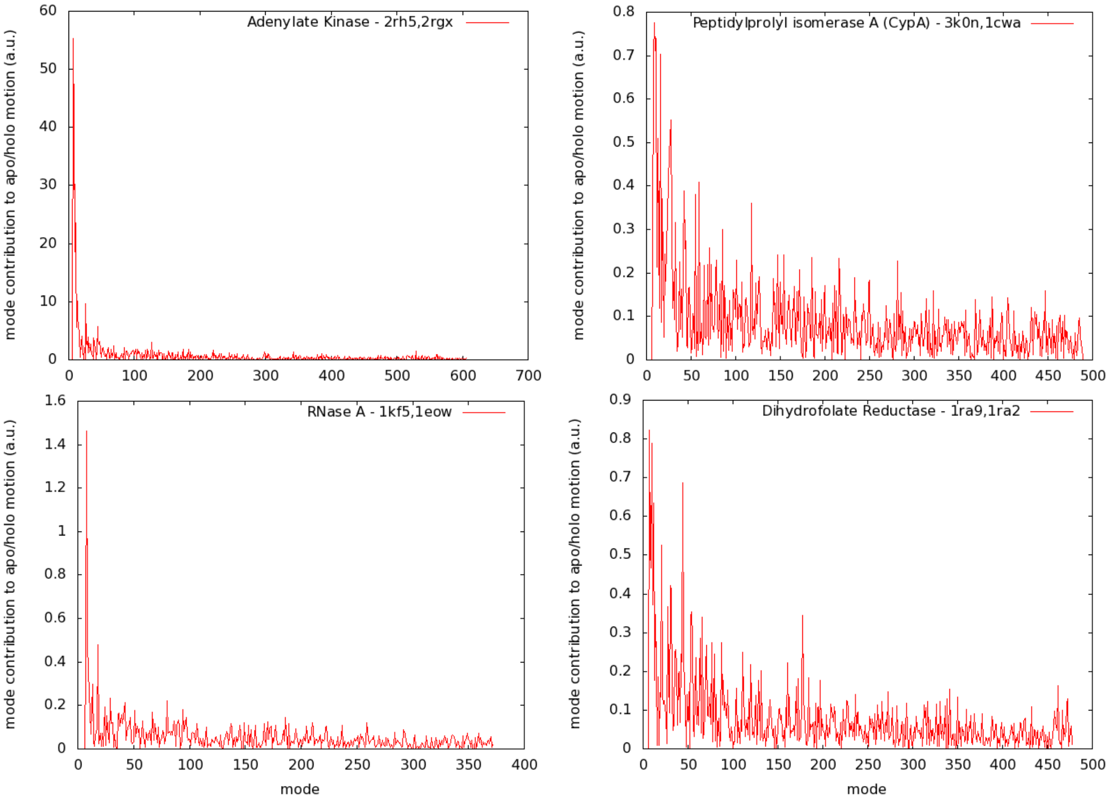


Figure 69 – Major contribution of the low energy normal modes to the motion between the free and bound form of the four model proteins known to bind their ligand by conformational selection.

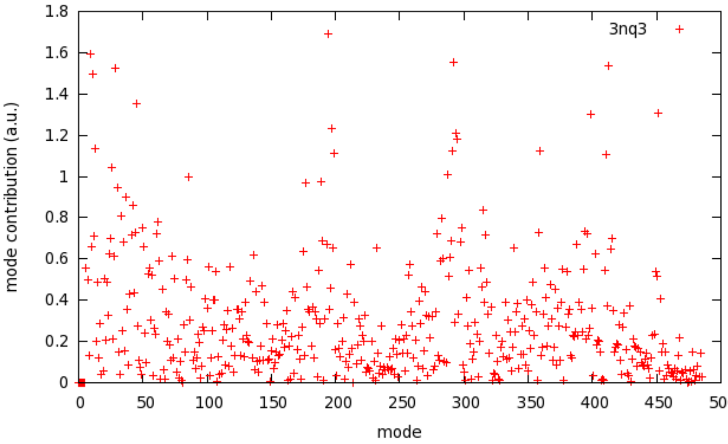


Figure 70 – Random contribution of each mode to the free/bound motion in beta-lactoglobulin; an enzyme with a buried hydrophobic ligand binding site and thus no option for ligand binding to occur by conformational selection.

While NMR CPMG relaxation dispersion data does not exist in the literature for T4L and HEWL, X-ray structures and NMR data does suggest that HEWL and T4L display conformational changes similar to the free/bound change in the absence of ligand. This is an indication of ligand binding by conformational selection. I therefore try to identify the binding site of T4L and HEWL with my NMA algorithm. The two proteins are classified as hinge bending proteins in the database of macromolecular motions.(Gerstein and Krebs 1998) Hinge motions are well described by normal modes. Thus I also expect the ligand binding algorithm to perform well on these two structures. Figure 71 shows the structure of T4L and HEWL with ligands in purple and surrounding surface probes colored in a gradient from blue across green to red. Those nodes that cause the largest change of equilibrium dynamics (i.e. has the lowest overlap with the mode 7 apo eigenvector) are colored in red. It can be seen that the most perturbing nodes all cluster around the active site.

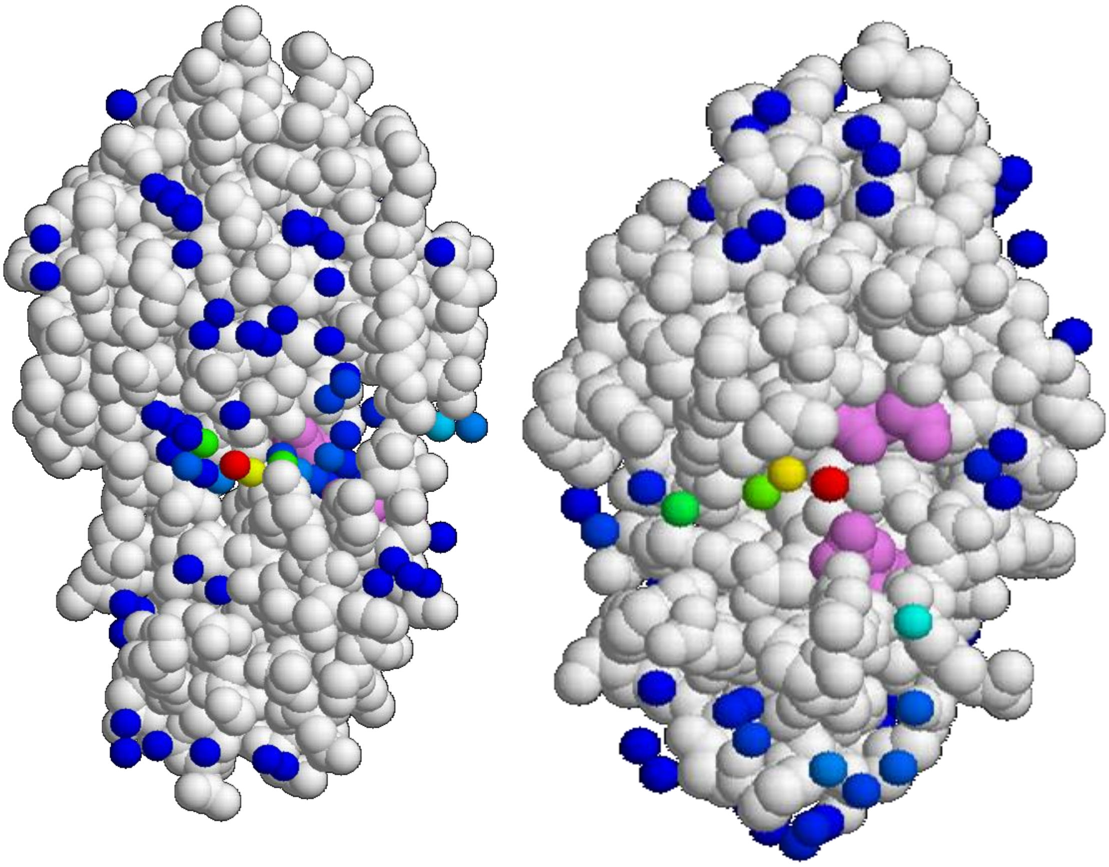


Figure 71 – T4L (left) and HEWL (right) shown as white space filling models. Catalytic residues are shown in violet. Probe atoms are colored from blue to red depending on their level of perturbation of the calculated normal modes. The probe atoms with the largest effect are all seen to be located in the ligand binding site cleft of the two proteins that both perform a hinge motion.

9.3.3 Identification of ligand binding sites in the larger dataset and comparison to the accuracy of other methods

I proceed to test my NMA algorithm for ligand binding site identification on a larger set of proteins. Success or failure of determining a ligand binding site is determined by whether the center of the predicted ligand binding site is within 6Å of the center of the center of the actual ligand binding site or not. The algorithms which my NMA method is benchmarked against are all described in section 9.1.5. The results for ligand binding site identification are summarized in Table 17 and Figure 72.

	All 99 proteins	21 hinge motion proteins	78 proteins without a hinge motion
POCASA	52	15	37
LIGSITE / ConCavity	57	16	41
My NMA method	16	14	2

Table 17 – The table summarizes the number of correct predictions of catalytic sites for a set of algorithms. The dataset consists of 99 ligand binding sites. The columns lists the number of successful predictions of catalytic sites for all proteins, hinge-motion protein and non-hinge motions proteins, respectively.

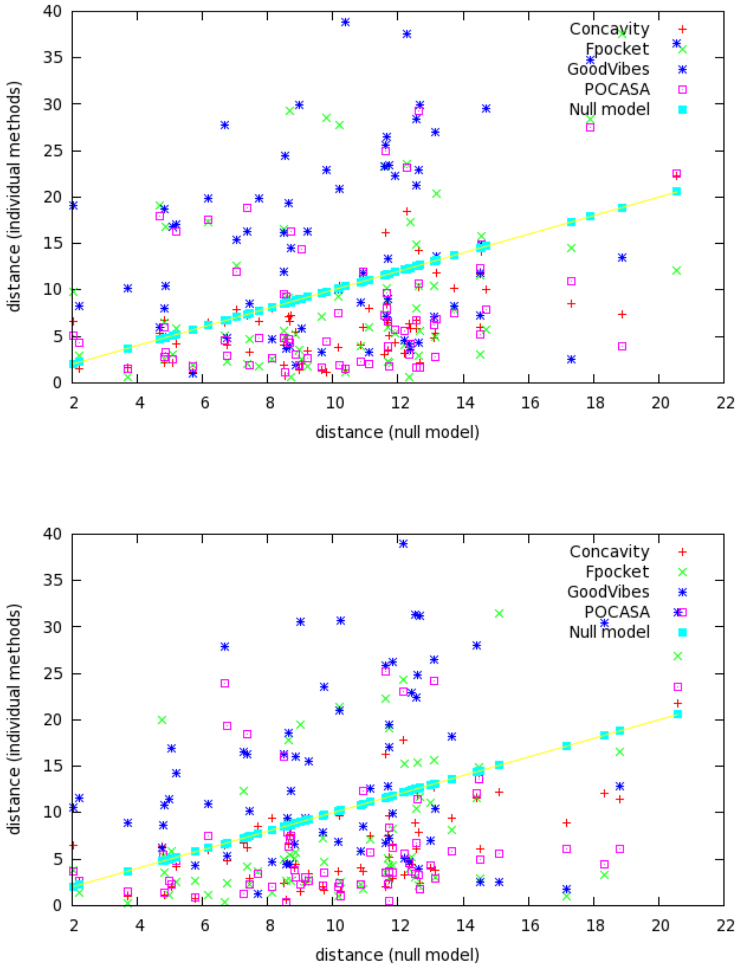


Figure 72 – Comparison of all methods based on the apo (top) and holo (bottom) structures. The null model is shown along the diagonal. Methods that perform better than the null method are plotted below the diagonal. My null model assumes the ligand binding site to be at the center of the protein.

There are examples of all methods being correct and all methods being wrong (e.g. 1i78A), but there are no examples of my NMA method making the correct prediction, while all other methods fails. Overall my NMA method performs worst of the benchmarked methods.

The 99 target structures can be divided into two subsets; one for which a hinge motion is observed and one for which a hinge motion is not observed. With success rates of 14 out of 21 on the first dataset and 2 out of 78 on the latter dataset it is clear that My NMA method performs much better, when a hinge motion is present. This is because hinge motions are often described by the normal mode with the lowest energy, and it the perturbation of the direction of normal mode 7, which is used for prediction of catalytic sites. There are no energy barriers preventing the bound and unbound state from being sampled. This in turn indicates that the ligand binding motion is sampled in the absence of the ligand, because there are no energy barriers between the bound and unbound state.

That my NMA method algorithm fails to predict the location of a large number of binding sites in cases where all the other algorithms succeed, might be because mode 7, which is the mode used for comparison before and after perturbation, does not always describe the motion between the bound and free state upon ligand binding. However, there is no clear indication that a high overlap is a guarantee of success (Figure 73). Figure 73 does not show that there is a correlation between the ability of my NMA method being able to predict ligand binding sites and whether the motion between free and bound structures are described by normal mode 7 or not. The higher the overlap between normal mode 7 and the vector between the free and bound structure, the better my NMA method does not perform; measured by the distance from the centre of the ligand to the centre of the predicted pocket.

Another explanation for the shortcomings of my NMA method might be that it predicts the binding site of allosteric inhibitors instead of the binding site of enzyme substrates. In the dataset I have however not found examples of enzymes with inhibitors bound in a site remote from the catalytic site, where the substrate is otherwise bound.

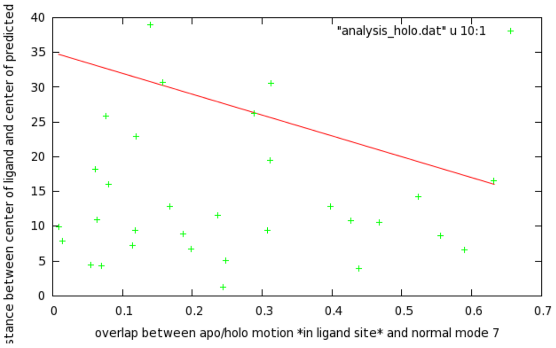


Figure 73 – Comparison of my NMA method performance - as measured by the distance between the center of the predicted ligand binding site and the center of the ligand - relative to the overlap between normal mode 7 and the vector between the apo and holo structure. Each dot represent an free/bound protein pair.

9.3.4 Residues in active sites are conserved and rigid

In this section I present a method using NMA whereby non-ligand binding residues can be identified. It is not a ligand binding site prediction method in itself, but it can be used to rule out false positives predicted by other algorithms. It has already been established that residues in active sites are conserved across species. (Capra and Singh 2007) Here I show that evolutionarily conserved residues are rigid due to dense packing as measured by B-factors and vice versa I show that flexible residues mutate more frequently. The magnitude and directionality of anisotropic B-factors have shown to correspond well with normal modes, which themselves describe concerted conformational changes of protein domains.(Atilgan, Durell et al. 2001; Bakan and Bahar 2009; Jackson, Foo et al. 2009) Therefore B-factors are excellent descriptors of protein dynamics. From the observation of the anti-correlation between sequence conservation and B-factors (Figure 74) one of two conclusions can be drawn. 1) A prerequisite for evolutionary mutation of residues is that those residues are positioned in flexible/dynamic regions of the protein. 2) A prerequisite for catalysis is that residues in the active site are pre-organized for electrostatic catalysis. As a consequence of the sequence conservation of rigid active site residues, other more flexible residues are free to mutate. The higher mutation rate of flexible residues is a consequence of them not being in the active site rather than them being flexible. Either way the conclusion is that residues in active sites are rigid and conserved. Rigid residues with low conservation scores (lower left corner of Figure 74) are observed, but flexible residues that are conserved are almost never observed (upper right corner).

I am aware that different residue types have different B-factor values (Table 18) and might have different degrees of sequence conservation. Because the anti-correlation observed in Figure 74 might be due to low B-factor residue types being conserved and vice versa I do the plot of B-factors against sequence conservation scores for each residue type (see appendix). The anti-correlation between the two properties describing flexibility and sequence conservation is observed for all 20 of the default amino acids with the exception of Cysteine (see appendix). This highlights the special role that Cysteine plays in proteins. It forms disulfide bridges and keeps the tertiary structure in check. A Cysteine participating in a disulfide bridge forms not 2 but 3 covalent bonds. This reduces the degrees of freedom for the residue and it most likely explains why high B-factors are not observed for Cysteine (Table 18) unlike its “oxygen counterpart” Serine, which displays above average B-factors. The disulfide bonding property of Cysteine is most likely also the explanation of its high sequence conservation score. This skews the data and no correlation is observed for Cysteine. Other residues with special properties are the hydrophobic residues (e.g. Val, Leu, Ile). They have smaller B-factors than other residue types due to their tight hydrophobic packing and frequent participation in rigid secondary structure elements. β-sheet residues on average have smaller B-factors than α-helix residues, which in turn have smaller B-factors than other residues. However the same anti-correlation between flexibility on one hand and sequence conservation on the other is observed for all residue types (see appendix).

From a viewpoint of thermostability it makes sense that rigid residues are conserved and flexible residues are free to mutate. Rigid residues like for example Proline decrease the configurational entropy of unfolding and thereby stabilize the folded protein.(Matthews, Nicholson et al. 1987) One would expect thermostable proteins to be favored by evolution. It has in fact been suggested that protein stability promotes evolvability.(Bloom, Labthavikul et al. 2006)

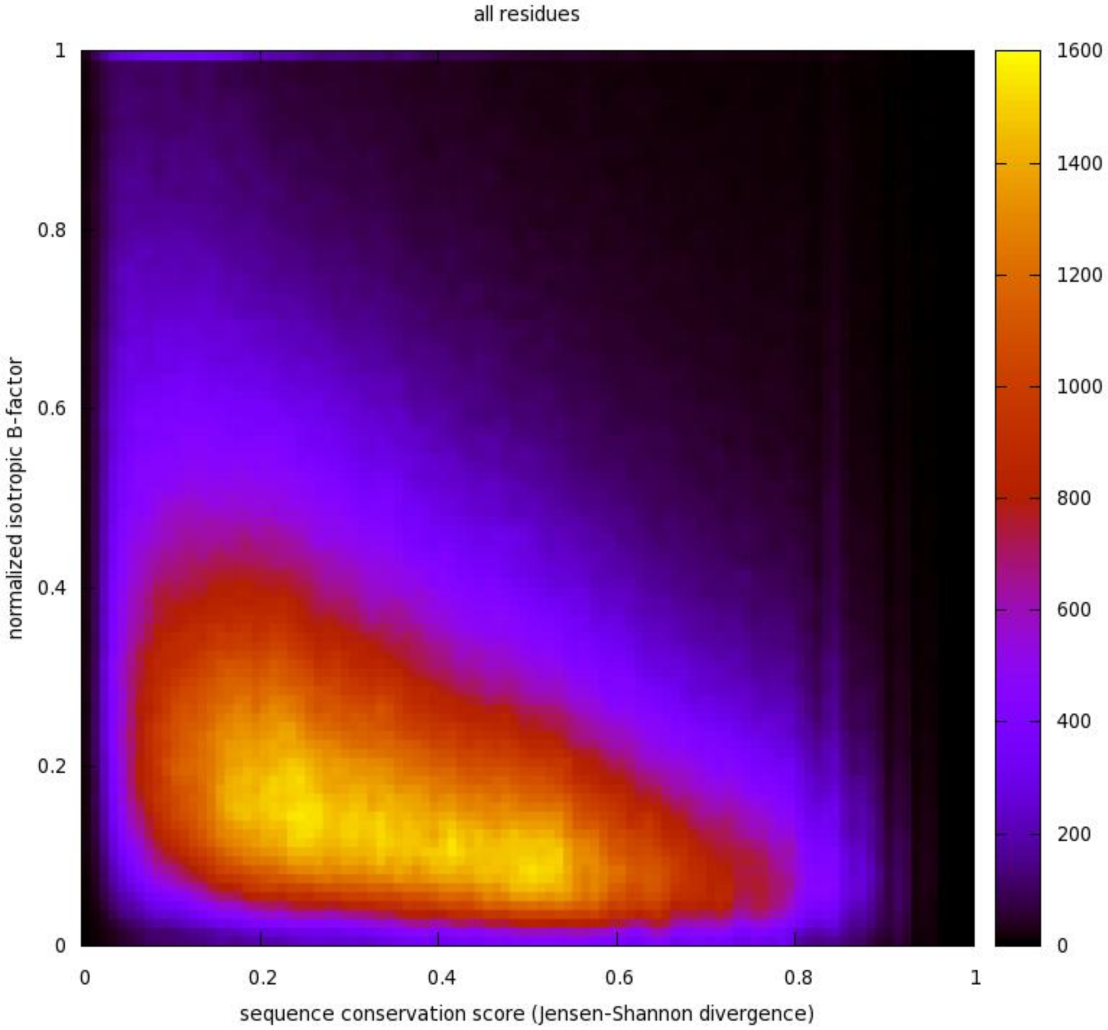


Figure 74 – Contour plot showing the anti-correlation between normalized isotropic B-factors (average of backbone B-factors for each residue) and sequence conservation scores (Jensen-Shannon divergence). The count of residues with each combination of the two parameters is shown on the third z-axis. The color coding of the contour plot is shown on the right. High sequence conservation scores are not associated with high B-factors, and there is a general anti-correlation (linear regression: r=-0.29, slope=-0.29) between B-factors and sequence conservation scores.

GLU	1.10	**GLN**	1.04	**ARG**	1.01	**MET**	0.97	**PHE**	0.93
LYS	1.09	**ASN**	1.04	**THR**	0.99	**LEU**	0.95	**ILE**	0.93
ASP	1.07	**SER**	1.03	**HIS**	0.98	**CYS**	0.95	**TYR**	0.93
PRO	1.04	**GLY**	1.02	**ALA**	0.97	**VAL**	0.93	**TRP**	0.92

Table 18 – Average normalized isotropic B-factors of backbone atoms for each of the 20 standard amino acid residues irrespective of secondary structure elements based on 3,340,239 residues in the PDB.

Having established that rigid residues are conserved and knowing that ligand binding residues are conserved, I turn to NMA and my four model proteins again. Specifically I plot the calculated residue fluctuations and mark the ligand binding residues (Figure 75). It can be seen that the ligand binding residues have a low probability of being flexible. Looking at residue flexibility, whether using NMA, X-ray, NMR or MD simulations, is therefore a method for ruling out false positives of ligand binding site finding algorithms. Applied to my own NMA algorithm I am able to rule out false positives using this method (Figure 76).

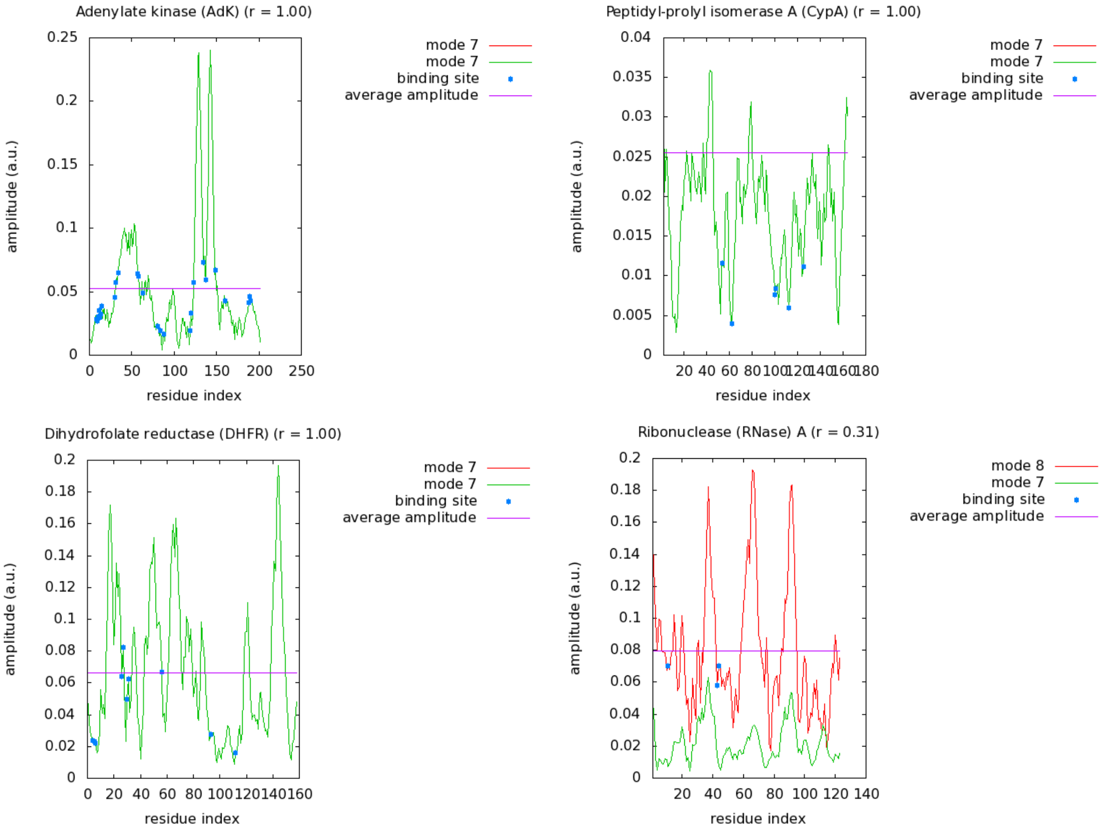
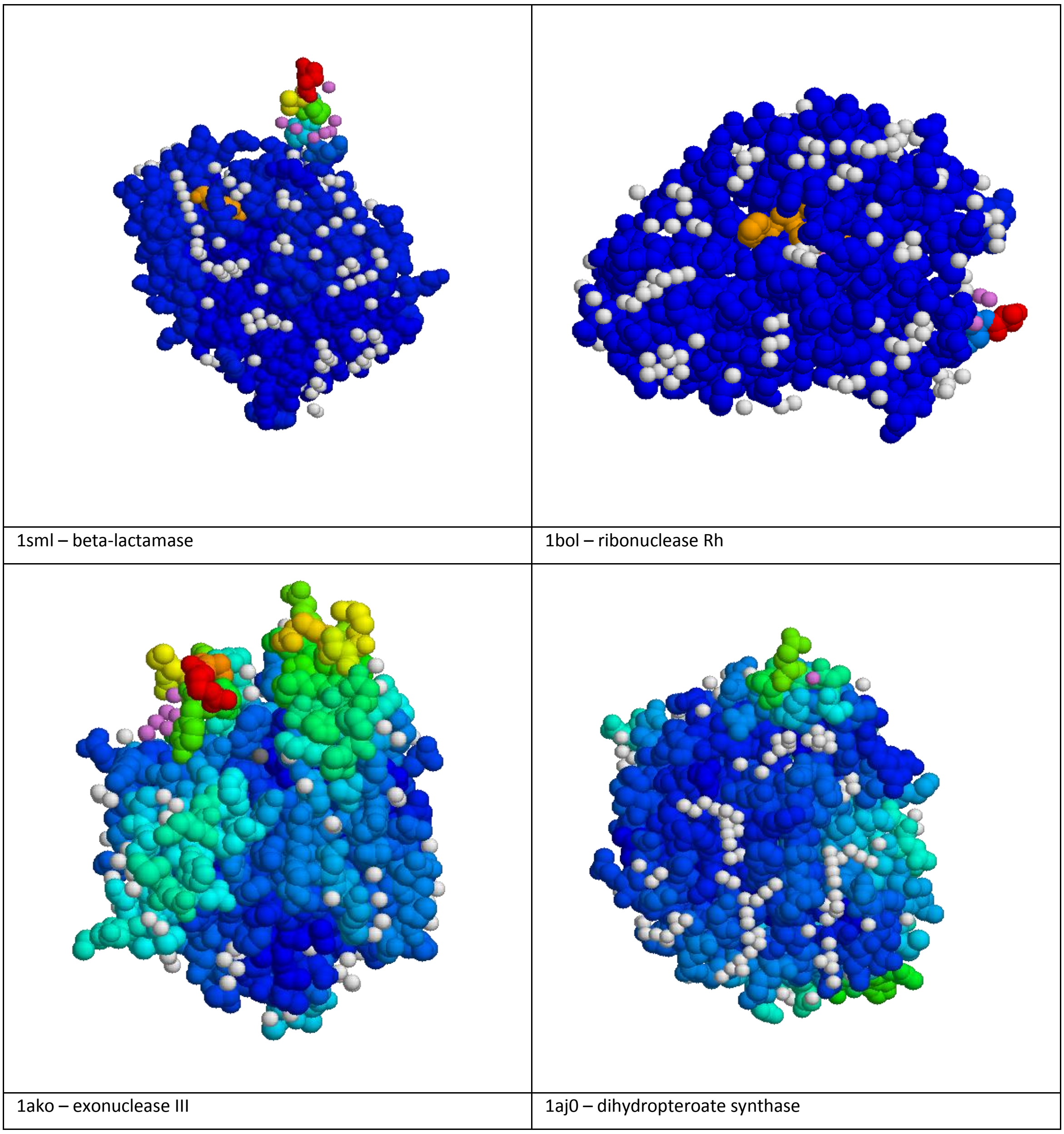


Figure 75 – Residue fluctuations of the four model proteins calculated with NMA. The mode with the highest free/bound overlap and normal mode 7 are shown. Ligand binding residues are displayed with blue dots. Ligand binding residues are in all cases located at local minima of residue fluctuations.



1sml – beta-lactamase

1bol – ribonuclease Rh

1ako – exonuclease III

1aj0 – dihydropteroate synthase

Figure 76 – Examples of predicted ligand binding sites located next to flexible residues. The grid points with the largest change in overlap are shown in violet. Other grid points are shown in white. The protein is colored from blue to red, with red residues being those displaying the largest amplitude according to normal mode 7.

9.4 Discussion

To direct site directed mutagenesis experiments it is important to having identified the ligand binding site and the catalytic residues. Despite research spanning decades and the number of structures growing each year the problem of ligand binding site prediction is still a challenge as evidenced by the accuracy of existing algorithms; the precision and recall of the prediction of catalytic residues is even lower.

I have presented a method for identification of ligand binding sites in proteins with ligand binding governed by conformational selection. In a benchmark against other methods I perform poorly. However, when identifying ligand binding sites in hinge bending proteins, the success rate of my method is comparable to other methods. This is an indication that NMA does not describe the functional motion in non-hinge bending proteins, normal mode 7 does not describe the functional motion and/or the proteins in the dataset do not bind their ligand by conformational selection, but rely heavily on induced fit. Another technical reason for the poor performance of my NMA algorithm might be because of the grid spacing (2Å), which is too coarse grained, and with the requirement of a minimum distance from the protein of (3Å) does not allow placement of grid points in buried and narrow ligand binding sites.

It has previously been shown that catalytic residues and inhibitor binding residues show little movement compared to other residues.(Yang and Bahar 2005) Instead it is the surrounding residues that move upon ligand binding in order to organize the ligand binding site. Therefore it did not seem unlikely that false positives predicted by my NMA method could be ruled out by ignoring grid points near residues with extreme fluctuations. These types of wrong predictions are one of the most frequent incorrect predictions made by my NMA method, so it is very relevant to eliminate them in order to approve the success rate of my NMA method. The problem is to automatically define whether a grid point is vicinal to a flexible residue or not. While I can visually identify the false positive grid points as being in the vicinity of flexible residues, it proved more difficult to write an algorithm that does not rule out true positives. Another problem is the definition of which residues are flexible. While I was unable to write an algorithm that automates the exclusion of grid points near flexible residues, I have presented cases involving visual inspection that confirm the absence of ligand binding sites near flexible residues.

Due to structural genomics initiatives the protein structure space is growing at a faster rate than ever. As these new proteins are characterized and their ligand binding site identified, ligand binding site prediction can in the future rely more on sequential comparison. Even without structures the growing number of genomes will enable a comparison to identify conserved residues that are obvious candidates for being ligand binding site residues.

9.5 Conclusion

It has been shown that NMA can be used to indentify ligand binding sites in proteins that bind their ligand by conformational selection. It has been shown that NMA can be used to identify residues that most likely do not interact directly with the bound ligand.

10 Discussion

The ability to engineer proteins is of great importance in the pharmaceutical and biotech industry. In this context it is important to understand the structural effect of mutations and how proteins interact with small molecule ligands and other proteins. To better understand and to develop algorithms for predicting the structural effect of mutations and ligand binding one needs to study a dataset and train the algorithm on this dataset. However, it is important to know, which conformational changes are real and which are due to intrinsic dynamics, experimental errors and external physiochemical properties. Therefore I chose to first do a thorough structural analysis of two proteins for which many mutant structures (T4L) and many structures at different experimental conditions (HEWL) have been solved. Most of these structures are solved by molecular replacement using a handful of structures as starting models. Conformations and even bad geometry is therefore inherited. To avoid restricting myself to a few starting models and two protein folds, I expanded the analysis to the entire PDB. The drawback of this is the requirement for full automation and the impossibility to analyze structures on an individual basis.

Understanding engineered proteins also means understanding how they interact with ligand. A first step in understanding ligand binding is the ability to be able to identify that ligand binding site in the first place. Energetic methods can be utilized, if the identity of the ligand is known. Otherwise one has to rely on geometric methods. Because of structural genomics initiatives the number of structures with unknown function and unknown ligand binding site is growing. Automated binding site prediction is therefore more relevant than ever in order to identify new binding sites and surfaces to get a better understanding of the energetic of these binding sites and how they can be engineered. In recent years multiple proteins have been suggested to bind their ligand by conformational selection. This implies there is no energy barrier between the ligand free and bound conformation and NMA should therefore be able to describe the motion. This spurred me to utilize NMA for ligand binding site prediction.

10.1 HEWL and T4L structural variability

I have shown that despite any motion being restricted by a great number of bond lengths, angles and dihedral angles, the variation between HEWL wt structures and T4L mutant structures is still quite significant. These different conformations yield slightly different electrostatic micro environments, which affect ligand binding and enzyme catalysis. From ensembles of X-ray structures and using NMA I have shown that the structural variability is not equally distributed across the structure. I have shown that the structural variability of each residue very much correlates with the B-factor of that residue. For both HEWL and T4L I have shown that the main cause of structural variation is not mutations, but rather crystal contact differences, when X-ray structures are solved in different space groups. I have also shown the importance of crystal contacts on the conformation of HEWL by showing the dependence of the conformation on the hydration level and thus the unit cell dimensions. I have shown that the reported conformation of HEWL and T4L is very dependent on especially the crystallographer and also the starting model in those cases where the phase problem is solved by molecular replacement. I have demonstrated that both rigid and flexible residues of HEWL are more identical, when comparing wt structures against their starting model than when comparing random wt structures against each other. I have revealed that the few sequence different mutant structures of HEWL in the PDB, which are solved from the same starting model, are more identical to each other and their starting model than random wt structures are to each other. I have shown that the structural variation in the two domains of HEWL is equally distributed in both the wt structures and the mutant structures and with the exception of the neighboring residues is independent of the location of the mutation. The location of the mutation in T4L affects the domain in which it is located more, and the structural variation of the small domain is larger, which NMA results also confirm. However, the structural effect of mutations seems to be independent of the distance from the site of mutation. This is an indication that mutations in T4L do not alter the overall conformation of T4L. The variation of solvent exposed and buried residues is similar, when comparing coordinate RMSDs, but side chain dihedrals are much more variable on the surface of the T4L than in its interior. I have presented a small number of cases, in which the HEWL structure is perturbed by small ions otherwise thought to be insignificant to the crystallographic structure. It has also been presented how polar ligands perturb the structure of a T4L mutant, whereas non-polar ligands do not change the conformation of the binding site residues. This is strong evidence that one needs to be careful that physiochemical conditions are identical, when comparing structures. I have however found no indication that pH and temperature differences perturb the conformational distribution of HEWL.

While conclusions about the structural effect of ligand binding and mutations are often drawn from single conformation X-ray structures, it would be preferable if these conclusions could be derived from ensembles of structures representing the conformational variation of a protein. For HEWL I have shown how the results of structure based calculations (pK_a values and stability changes) rely heavily on the structure fed to the program. Interestingly the experimental result is always sampled among the different conformations. This is testimony to the importance of carrying out structure based calculations on ensembles of structures instead of individual structures that represent end states and carry experimental errors. Despite crystallographers being presented with electron density from two alternate conformations, only the conformation of the most prevalent conformation is usually given. As more high resolution structures are solved, this will undoubtedly reveal a lot about the conformational variability of proteins.

10.2 Protein structure variation in the entire PDB

I have analyzed the conformational variation in the entire PDB. Like I found for HEWL and T4L it is the space group, which is the most important parameter for structural variation. In fact I did not find a single pair of space groups with structures which had a lower RMSD between them than structures from a single space group. I also found that the contraction/expansion of the unit cell and the concomitant change of hydration level, Matthews coefficient and ultimately crystal contacts had an almost linear relationship with RMSD. The Matthews coefficient is tied to the crystallographic resolution, but even when taking this into account, there is still an effect on protein conformation by the Matthews coefficient and thus the crystal contacts. Because crystal contacts are so important, it is not entirely accurate to talk about intrinsic dynamics when observing conformational differences between protein with different crystal contacts.(Henzler-Wildman, Thai et al. 2007) Although crystal contacts dominate the protein conformation I also found an effect of pH, temperature, ions, author/crystallographer and structure quality as measured by resolution and the presence/absence of missing residues. While the correlation between RMSD and these parameters is not linear, they all nevertheless had an effect on the reported conformations. All these parameters are in the end tied to the crystallographer. It should be a big concern that specific conformations of proteins are so dependent on the author. I looked for statistical interactions between author and physiochemical parameters to identify, whether the physiochemical parameters also had an effect in cases of identical authors. A crystallographer however will often use identical conditions, and the statistical test was weakened by the population asymmetry and no conclusions could be drawn.

While I did find mutations to have an effect on the overall conformation of proteins, I did not find proteins differing by only one residue to have different conformations. On the contrary single point mutants had lower RMSDs between them than wild type structures, which once again is a testament to the importance of molecular replacement on the final reported conformation. I also investigated the dependence of conformational variation on the distance from the site of mutation in the case of single mutations. I did however find no apparent effect of a single mutation neither in the vicinity nor distant from the site of mutation. Mutations do not alter the overall conformation of T4L and other proteins in general. Therefore the focus when modeling mutations in silico should be entirely on electrostatics of the local environment. It is to a large degree the surrounding residues, which determine the backbone conformation of a residue (appendix). Any changes due to a mutation will therefore only have to be modeled for side chain conformations in the absolute vicinity of the mutation, if steric clashes or unfavorable electrostatics are introduced.

10.3 Prediction of ligand binding sites with NMA

I have developed an NMA algorithm for predicting the anisotropic motion of proteins at equilibrium. I have validated my NMA algorithm by comparison of the calculated isotropic amplitudes to isotropic B-factors and apparent fluctuations between NMR ensembles. The correlation of calculated amplitudes with X-ray temperature factors validates the use of NMA for description of the amplitude of movement. I have shown that the calculated eigenvectors do not depend on the input structure in more than 99.9% of all cases. This is an indication that motions at equilibrium are a property built into the structure and independent of perturbation of the structure by the presence of ligands or single point mutations. No gold standard exists for choosing a proper cutoff distance for the elastic network of NMA calculations. Since the cutoff distance is the single most important parameter in an elastic network model involving uniform force constant, it is important to verify that similar results can be obtained within a range of cutoff distances. I have shown the NMA calculations to be reproducible for T4L over a wide range of cutoffs. I have shown that NMA can be used to describe conformational changes between open and closed states of proteins. That a few eigenvectors describe the motion between an open and closed conformation validates the use of NMA for description of the direction of movement. While not successful for all proteins, it has proven to be successful for all proteins known to bind their ligand by conformational selection. For these same four proteins binding their ligand by conformational selection I have shown that NMA can be used to predict ligand binding sites by perturbing the elastic network in a grid based fashion and recalculating eigenvectors and compare those eigenvectors to the non-perturbed system. I have further established a link between flexibility and residue conservation and used the fact that ligand binding residues are conserved to develop a method for excluding residues as ligand binding residues, if they are flexible and thus unlikely to be conserved. This has proven to be a useful method for hinge motion proteins in which the ligand binding site was otherwise incorrectly predicted.

The motions calculated by all-atom detailed force field NMA correlate with the motions calculated by alpha-carbon simple force field NMA(Tirion 1996), which validates the use of fast coarse-grained GNM methods for description of functionally important motions over slower classical NMA. That conformational changes can be addressed with normal mode analysis also makes the method superior to molecular dynamics in terms of speed.

A disadvantage of NMA is that the calculated trajectories are linear and a linear pathway between two conformations might be sterically impossible. An algorithm has been developed by Thorpe and coworkers, which sample the sterically allowed conformational space.(Wells, Menor et al. 2005) Applying the methods of FRODA to the NMA binding site prediction algorithm might make the algorithm even more powerful in terms of describing conformational changes prior to ligand interaction.

One reason I do not always succeed in finding binding sites with my NMA algorithm, might be because NMA does not succeed in describing the relevant functional motion. To validate the motions predicted by normal mode analysis could perhaps be done by measuring residual dipolar couplings (RDCs), which yields angles and distances between atoms during fluctuation. This would serve to validate the coordinate changes predicted by NMA. According to Akke(Akke 2002) it should be possible to apply this method. The motions calculated with NMA should also be compared to molecular dynamic (MD) simulations. MD simulations will yield not only an end conformation but also a trajectory. The results from NMA can be compared to the conformations along this trajectory.

Ligand binding by conformational selection is still an emerging field and few mutational studies exist. To better understand the connection between ligand binding, catalytic turnover and the amplitude and timescale of intrinsic dynamics, more measurements on more model proteins and more mutants must be carried out. It would be interesting upon mutation to measure changes in both activity as measured by calorimetry and timescales as measured by NMR relaxation and fluctuation directions and amplitudes as measured by residual dipolar couplings (RDCs).

I have observed that NMA in general reproduce hinge motions better than shear motions(Krebs, Alexandrov et al. 2002) from the database of molecular movements(Flores, Echols et al. 2006). This is an indication that either NMA is more robust at predicting conformational changes of hinge bending proteins or hinge bending proteins are designed to bind their ligand by conformational selection, while other proteins rely on binding their ligand by induced fit. Therefore I believe that the search for new model proteins binding their ligand by conformational selection should be carried out among hinge bending proteins. One such protein in which the hinge bending motion has been shown to be linked to substrate binding and catalytic turnover is thermolysin.(Veltman, Eijsink et al. 1998)

The use of NMA is not limited to predicting flexibility and ligand binding sites. Others have used NMA for introducing flexibility in protein-protein docking(Zacharias and Sklenar 1999; Cavasotto, Kovacs et al. 2005; Lindahl and Delarue 2005; Mitra, Schaffitzel et al. 2005; Dobbins, Lesk et al. 2008; May and Zacharias 2008). The method is primarily evaluated by the ability to reduce the root mean square deviation between the coordinates of the docked proteins and the coordinates obtained from the normal mode trajectory. If it could be possible to get closer to the structure in the interaction complex before carrying out protein-protein docking it would inarguably be of tremendous value. Other uses of NMA that I have thought of include the following. 1) Generation of ensembles of structures for use in molecular replacement. 2) Determination of non-flexible parts of a structure, which can be used for NCS restraints during crystallographic model refinement (Rupp 2009), p. 634). This method is only applicable if differences between identical NCS related molecules is due to intrinsic dynamics that can be described by NMA rather than being due to crystal contact differences. 3) Weighting of residues by dynamic properties during sequential and especially structural alignment. Non-flexible residues should have a higher importance; i.e. those residues should score higher than other residues if aligned with each other.

11 Conclusion

I have shown that the reported conformations of proteins in the Protein Data Bank are highly dependent on the starting model used for molecular replacement. It can therefore be concluded that when analyzing the effect of mutations and ligand binding, it is only valid to compare structures against their starting models. I have shown that single point mutations in general have no apparent effect on the overall structure of proteins. I have shown that intrinsic dynamics and/or experimental error contribute more to conformational variation than mutations. I have shown that the results of structure based calculations are highly dependent on the conformation of the input structure. When performing structure based calculations such as pK_a calculations, binding energy calculations and stability change ($\Delta\Delta G_{\text{mutation}}$) calculations those calculations should be carried out on an ensemble of conformations that sample all the states, which a protein occupies in solution. Carrying out a calculation on just one conformation of a protein will not reveal the range of values, which are attained at the dynamic equilibrium. Especially in the case of protein-protein interaction energy calculations NMA could be such a method for generating an ensemble of structures, because the interactions of two protein surfaces involve large scale concerted conformational changes.

I have demonstrated that normal mode analysis can be used to determine ligand binding sites in proteins binding their ligand by conformational selection and in proteins with hinge motions.

12 Abbreviations

12.1 Metrics

GDT – Global Distance Test

RMSD – Root Mean Square Deviation

RMSF – Root Mean Square Fluctuation

12.2 X-ray crystallography

ASU – Assymetric Unit

MAD – Multi-wavelength Anomalous Dispersion

MIR – Multiple Isomorphous Replacement

MIRAS – MIR with Anomalous Scattering

MR – Molecular Replacement

NCS – Non-Crystallographic Symmetry

SAD – Single-wavelength Anomalous Dispersion

SIR – Single Isomorphous Replacement

SIRAS – SIR with Anomalous Scattering

12.3 Molecules

AdK – adenylate kinase

BME – beta-mercapto ethanol

CypA – peptidylprolyl isomerase A

DHFR – dihydrofolate reductase

HEWL – hen egg white lysozyme

NADPH – nicotinamide adenine dinucleotide phosphate

RNase A – ribonuclease A

T4L – bacteriophage T4 lysozyme

12.4 Normal Mode Analysis

ANM – Anisotropic Network Model

ENM – Elastic Network Model

GNM - Gaussian Network Model

NMA – Normal Mode Analysis

12.5 Other

BMRB - Biological Magnetic Resonance Bank(Ulrich, Akutsu et al. 2008)

CASP - Critical Assesment of Structure Prediction

CPK – Corey, Pauling, Koltun

CPMG - Carr-Purcell-Meiboom-Gill

KIE – Kinetic Isotope Effect

MD – Molecular Dynamics

mmCIF – macromolecular Crystallographic Information File

NMR – Nuclear Magnetic Resonance

PCR – Polymerase Chain Reaction

PDB - Protein Data Bank

vdW – van der Waals

13 Appendix

13.1 Protein structure variation

13.1.1 T4L structures in the PDB

[www.proteinkemi.dk/table_T4L.txt]

Table 19 – List of 392 HEWL structures used for analysis.

13.1.2 HEWL structures in the PDB

[www.proteinkemi.dk/table_HEWL.txt]

Table 20 – List of 172 T4L structures used for analysis.

13.1.3 Biological unit transformations

If matrix transformations are applied to get from the asymmetric unit to the biological unit, then the identity of the biological units being compared are very dependent on the correct determination of the dimensions of the asymmetric unit and the position of the protein within the asymmetric unit. I have shown the RMSD to be higher when transformations are applied ($\langle RMSD \rangle_{\text{transformations}}$ = .89Å) relative to when they are not applied ($\langle RMSD \rangle_{\text{no transformations}}$ = .61Å). However, this difference is only significant, when considering structures from identical and different space groups.

Transformations will at least double the number of chains. If there is a correlation between the number of chains and the RMSD, then the conclusion about transformations having an effect on the overall RMSD should be re-evaluated.

13.1.4 Number of chains

The RMSD is dependent on the number of chains but not in a linear fashion. This can be seen by plotting the RMSD as a function of the number of chains (Figure 39). This I do for the structure pairs that have identical space groups and 0 mutations, because it is expected that there is a linear correlation between the number of mutations and the average RMSD. The data is heavily skewed, as most proteins in the dataset are monomeric, and there is no linear correlation (r = 0.08). Therefore any slope determined by linear regression should be scrutinized. However it seems apparent that a different number of chains cause a different average RMSD (Figure 39). Therefore I choose to only focus on monomeric proteins from this point forward.

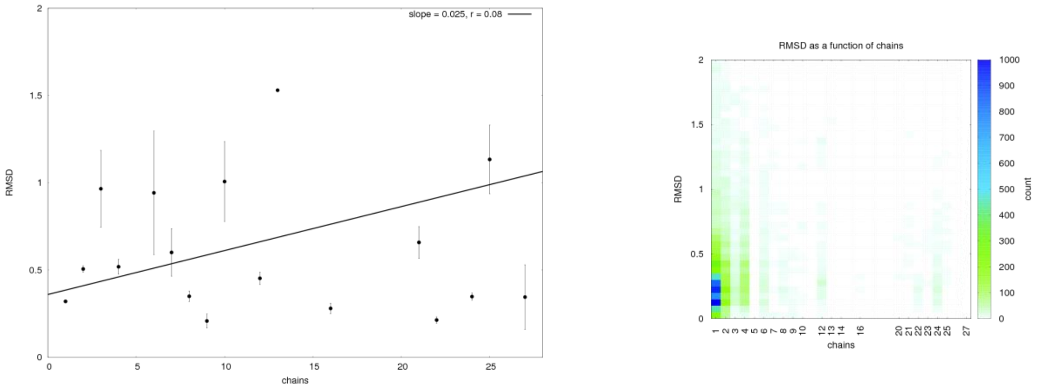


Figure 77 – RMSD on the y-axis plotted against the number of chains on the x-axis. The RMSDs vary, but there is no linear relationship between the number of chains and the RMSD (*n* = 9976, *r* = 0.08, $n_{averages}$ = 21, $r_{averages}$ = 0.02, $p(r_{averages}=0)$ = 0.94).

13.1.5 Resolution

It is important to remember that the reported resolution of a structure, d_{min}, is just the high end of the resolution range. The reported resolution is not an exact measure of structure quality. The number should be adjusted for the crystallographic *R*-factor (and the free R-factor), which is a measure of the observed electron density not accounted for by modeled atoms. Alternatively the use of an effective resolution(Weiss 2001), d_{eff}, could be used, but this requires knowledge of the completeness, *C*. Whereas the resolution, d_{min}, is always available from structure files, the completeness, *C*, is not always reported in PDB/mmCIF files, which makes the calculation of the effective resolution, d_{eff}, unfeasible.

$$d_{eff} = d_{\min} C^{-1/3} \qquad 13\text{-}1$$

13.1.6 Mutation from Glycine to Alanine

Is the conformation of a residue pre determined by its position in the protein or is the conformation of a residue an intrinsic property of that residue? Here I present evidence for the former. Mutating Glycine to Alanine will in most cases restrict the Ramachandran angles of those Alanine residues to an area usually only occupied by Glycine; i.e. 50 < ϕ < 100, 0 < ψ 50.

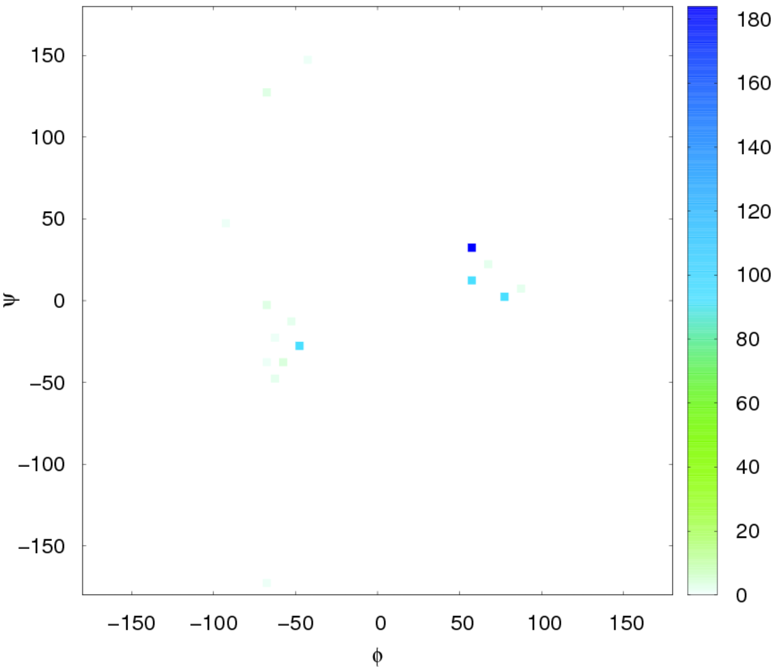
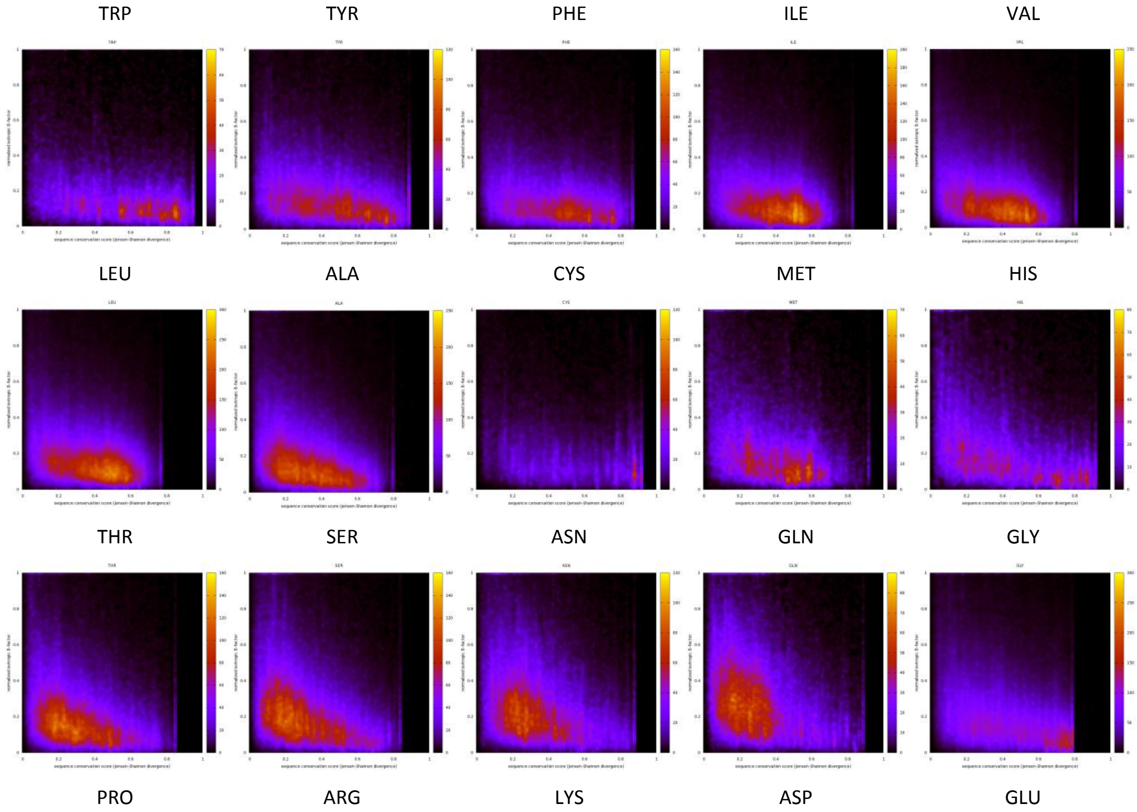


Figure 78 - ϕ,ψ-angles of alanine residues upon mutation from glycine. The Ramachandran angles of the alanine residues all map to a region usually only accommodated by glycine.

13.2 Ligand binding site identification

13.2.1 Residues in active sites are conserved and rigid



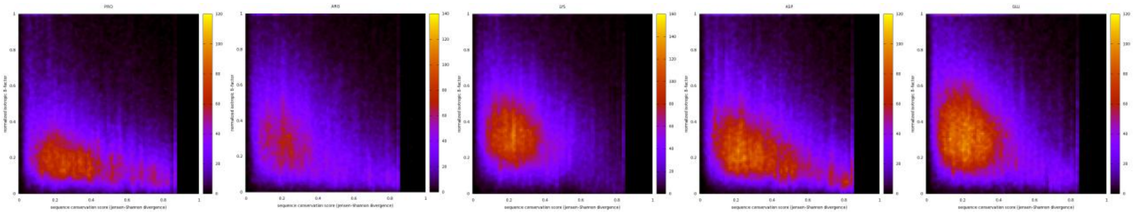


Figure 79 – Plot of B-factor as a function of sequence conservation score for each of the 20 standard amino acid residues. The hydrophobic generally non-flexible amino acid residues are shown at the top of the plot, and the charged and more flexible residues are shown at the bottom of the plot.

13.2.2 Induced fit and conformational selection

There is plenty of structural evidence for conformational differences in ligand free and bound states in the Protein Data Bank (PDB).(Rini, Schulze-Gahmen et al. 1992; M. Jack Borrok 2007; Aspeslagh, Li et al. 2011; Kadirvelraj, Sennett et al. 2011; Koch, Heine et al. 2011; Petty, Emamzadah et al. 2011) However, it is less well understood which role conformational changes play in ligand binding.

The conformational change upon ligand binding can be caused by a combination of the intrinsic motions of the protein and the structural interaction of protein and ligand. One can view ligand induced transitions as a transition between Boltzmann distribution of states.(Wrabl, Gu et al. 2011) A conformational change observed upon ligand binding is then the result of a combination of conformational selection and induced fit.(Niu, Bruschweiler-Li et al. 2011; Silva, Bowman et al. 2011) It is important to stress that conformational selection and induced fit are not mutually exclusive.(Hammes, Chang et al. 2009; Silva, Bowman et al. 2011)

13.2.3 Experimental evidence of conformational selection

The experimental evidence for conformational selection is growing. More methods are being utilized and conformational selection is being demonstrated in a growing number of proteins. The methods for showing ligand binding by conformational selection include NMR CPMG relaxation dispersion experiments, high resolution X-ray crystallography, binding studies above and below the “glass” transition temperature, studies of the effect on conformational change of mutations distant from the binding site and small-angle X-ray solution scattering in combination with coarse grained computational simulations.(Yang, Blachowicz et al. 2010)

One method of indicating conformational selection is to show that the rate of the conformational exchange (k_{ex}) measured by for example NMR CPMG relaxation dispersion and the catalytic rate (k_{cat}) are on the same time scale. This would indicate ligand binding or release to be the rate limiting step and thus ligand binding to be dependent on a conformational change. This has been demonstrated for CypA(Eisenmesser, Millet et al. 2005), RNase A(Cole and Loria 2002; Beach, Cole et al. 2005), DHFR(Falzone, Wright et al. 1994; McElheny, Schnell et al. 2005; Boehr, McElheny et al. 2006), AdK(Wolf-Watz, Thai et al. 2004; Henzler-Wildman, Lei et al. 2007; Henzler-Wildman, Thai et al. 2007) and triosephosphate isomerase.(Williams and McDermott 1995) In AdK, CypA and RNase A the time scale of motion is independent of the presence of ligand, whereas the time scale of motion is dependent on the presence of ligand in DHFR.

A second method of indicating ligand binding by conformational selection is to show that NMR relaxation dispersion chemical shift changes ($\Delta\omega$) at equilibrium in the absence of ligand correlate with chemical shift changes ($\Delta\delta$) upon ligand binding. This has been shown for all four of my model proteins.(Wolf-Watz, Thai et al. 2004; Beach, Cole et al. 2005; McElheny, Schnell et al. 2005; Boehr, McElheny et al. 2006)

Mutational evidence for the importance of conformational flexibility to enzyme catalysis involves showing that a mutation of a residue not interacting with the ligand disrupts the conformational flexibility and reduces the catalytic activity. This has been shown by NMR for RNase A(Kovrigin and Loria 2006) and DHFR(Bhabha, Lee et al. 2011). Furthermore point mutations in a phosphotriesterase positioned distant from the active site, which change the conformational populations of the enzyme but not the active site itself, have been shown to change the catalytic turnover of the enzyme.(Jackson, Foo et al. 2009)

Other NMR methods and methods other than NMR exist for studying conformational changes of proteins. One of them is X-ray crystallography. X-ray structures in the PDB provide many examples of structural differences between proteins with and without bound ligands, but the static structures at the end points do not reveal, whether the conformational change happened due to ligand binding (induced fit) or whether the ligand was able to bind, because the protein changed conformation at a pre-existing equilibrium (conformational selection). However, for AdK it has been shown that multiple conformations from the same asymmetric unit lie on the trajectory between a fully open and fully closed structure.(Henzler-Wildman, Thai et al. 2007) In the case of AdK the loop covering the active site displays a large correlated motion when the enzyme is isolated in solution in the absence of substrate. It turns out that this motion is very similar to the conformational change observed upon ligand binding. (Figure 54). Energy calculations also confirm that no barrier separates the ligand free and bound conformation of AdK.(Arora and Brooks 2007) Thus AdK presents a case where the conformational change necessary for function has been encoded in the protein structure, and the protein undergoes this conformational change, even when there is no substrate around. By mutational stabilization of a hidden conformation in CypA, it has been possible to observe the less populated conformation of CypA by X-ray crystallography.(Fraser, Clarkson et al. 2009) For RNase A it has been shown by high resolution X-ray crystallography that substrate and an inhibitor is bound at 228K but not at 212K and once bound the inhibitor cannot be washed off below 212K.(Rasmussen, Stock et al. 1992) This is an example of a “glass” transition temperature(Vitkup, Ringe et al. 2000; Teeter, Yamano et al. 2001) and it suggests the importance of conformational selection to ligand binding in RNase A. It has also been shown that mutations that shift the population between the open and closed conformation of a multi domain protein without affecting the binding site also modulate the binding affinity of that protein, which is an indication of ligand binding by conformational selection.(Mackereth, Madl et al. 2011) And likewise it has been shown by X-ray crystallography for protein kinase A, that the ligand free structure is represented by an ensemble of ligand bound and ligand free conformations.(Badireddy, Yunfeng et al. 2010; Masterson, Cheng et al. 2010) Likewise NMR has provided structural evidence that bound conformations are sampled in the absence of ligand.(Lange, Lakomek et al. 2008; Anthis, Doucleff et al. 2011)

Through the combination of low resolution small-angle X-ray solution scattering (SAXS) and coarse grained computational simulations it has been shown for a multi domain tyrosine kinase that in the absence and presence of a ligand the same bound and unbound states are populated.(Yang, Blachowicz et al. 2010) The ligand does not induce binding. Instead it is the population size of each state, which is changed upon ligand binding. This is evidence of the sampling of the unbound and bound state being completely independent of a ligand. Instead of a ligand-induced fit, multiple conformational states are sampled at equilibrium in solution; among them the bound state.

14 Bibliography

Abdul Ajees, A., K. Gunasekaran, et al. (2006). "The structure of complement C3b provides insights into complement activation and regulation." Nature **444**(7116): 221-225.

Aggarwal, A. K., D. W. Rodgers, et al. (1988). "Recognition of a DNA operator by the repressor of phage 434: a view at high resolution." Science **242**(4880): 899-907.

Ahmed, A., F. Rippmann, et al. (2011). "A Normal Mode-Based Geometric Simulation Approach for Exploring Biologically Relevant Conformational Transitions in Proteins." Journal of Chemical Information and Modeling: null-null.

Ahmed, A., S. Villinger, et al. (2010). "Large-scale comparison of protein essential dynamics from molecular dynamics simulations and coarse-grained normal mode analyses." Proteins: Structure, Function, and Bioinformatics: n/a-n/a.

Ajees, A. A., G. M. Anantharamaiah, et al. (2006). "Crystal structure of human apolipoprotein A-I: Insights into its protective effect against cardiovascular diseases." Proceedings of the National Academy of Sciences of the United States of America **103**(7): 2126-2131.

Akke, M. (2002). "NMR methods for characterizing microsecond to millisecond dynamics in recognition and catalysis." Curr Opin Struct Biol **12**(5): 642-7.

Alber, T., S. Dao-pin, et al. (1987). "Contributions of hydrogen bonds of Thr 157 to the thermodynamic stability of phage T4 lysozyme." Nature **330**(6143): 41-46.

Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." Journal of Molecular Biology **215**(3): 403-410.

Anthis, N. J., M. Doucleff, et al. (2011). "Transient, Sparsely Populated Compact States of Apo and Calcium-Loaded Calmodulin Probed by Paramagnetic Relaxation Enhancement: Interplay of Conformational Selection and Induced Fit." Journal of the American Chemical Society **133**(46): 18966-18974.

Apweiler, R., A. Bairoch, et al. (2004). "UniProt: the Universal Protein knowledgebase." Nucleic Acids Research **32**(suppl 1): D115-D119.

Arnold, G. E. and R. L. Ornstein (1992). "A molecular dynamics simulation of bacteriophage T4 lysozyme." Protein Eng. **5**(7): 703-714.

Arora, K. and C. L. Brooks (2007). "Large-scale allosteric conformational transitions of adenylate kinase appear to involve a population-shift mechanism." Proceedings of the National Academy of Sciences **104**(47): 18496-18501.

Artymiuk, P. J., C. C. F. Blake, et al. (1979). "Crystallographic studies of the dynamic properties of lysozyme." Nature **280**(5723): 563-568.

Ashkenazy, H., R. Unger, et al. (2011). "Hidden conformations in protein structures." Bioinformatics **27**(14): 1941-1947.

Aspeslagh, S., Y. Li, et al. (2011). "Galactose-modified iNKT cell agonists stabilized by an induced fit of CD1d prevent tumour metastasis." EMBO J **30**(11): 2294-2305.

Atilgan, A. R., S. R. Durell, et al. (2001). "Anisotropy of Fluctuation Dynamics of Proteins with an Elastic Network Model." Biophys. J. **80**(1): 505-515.

Atkins, P. W. and J. de Paula (2002). Atkins' Physical Chemistry.

Baase, W. A., L. Liu, et al. (2010). "Lessons from the lysozyme of phage T4." Protein Science **19**(4): 631-641.

Badireddy, S., G. Yunfeng, et al. (2010). "Cyclic AMP analog blocks kinase activation by stabilizing inactive conformation: Conformational selection highlights a new concept in allosteric inhibitor design." Molecular & Cellular Proteomics.

Bahar, I., A. R. Atilgan, et al. (1997). "Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential." Folding and Design **2**(3): 173-181.

Bahar, I. and Q. Cui (2005). Normal Mode Analysis: Theory and Application to Biological and Chemical Systems.

Bakan, A. and I. Bahar (2009). "The intrinsic dynamics of enzymes plays a dominant role in determining the structural changes induced upon inhibitor binding." Proc Natl Acad Sci U S A **106**(34): 14349-54.

Bava, K. A., M. M. Gromiha, et al. (2004). "ProTherm, version 4.0: thermodynamic database for proteins and mutants." Nucl. Acids Res. **32**(suppl_1): D120-121.

Beach, H., R. Cole, et al. (2005). "Conservation of Î¼sâˆ'ms Enzyme Motions in the Apo- and Substrate-Mimicked State." Journal of the American Chemical Society **127**(25): 9167-9176.

Berman, H. M., J. Westbrook, et al. (2000). "The Protein Data Bank." Nucl Acids Res **28**: 235 - 242.

Bernal, J. D., I. Fankuchen, et al. (1938). "An X-Ray Study of Chymotrypsin and Hemoglobin." Nature **141**(3568): 523-524.

Bernstein, F. C., T. F. Koetzle, et al. (1977). "The protein data bank: A computer-based archival file for macromolecular structures." Journal of Molecular Biology **112**(3): 535-542.

Best, R. B., K. Lindorff-Larsen, et al. (2006). "Relation between native ensembles and experimental structures of proteins." Proceedings of the National Academy of Sciences **103**(29): 10901-10906.

Betancourt, M. R. and J. Skolnick (2001). "Universal similarity measure for comparing protein structures." Biopolymers **59**(5): 305-309.

Bhabha, G., J. Lee, et al. (2011). "A Dynamic Knockout Reveals That Conformational Fluctuations Influence the Chemical Step of Enzyme Catalysis." Science **332**(6026): 234-238.

Birdsall, B., J. Feeney, et al. (1980). "The use of saturation transfer NMR experiments to monitor the conformational selection accompanying ligandâ€"protein interactions." FEBS Letters **120**(1): 107-109.

Blake, C. C. F., D. F. Koenig, et al. (1965). "Structure of Hen Egg-White Lysozyme: A Three-dimensional Fourier Synthesis at 2 [angst] Resolution." Nature **206**(4986): 757-761.

Bloom, J. D., S. T. Labthavikul, et al. (2006). "Protein stability promotes evolvability." Proceedings of the National Academy of Sciences **103**(15): 5869-5874.

Boehr, D. D., D. McElheny, et al. (2006). "The Dynamic Energy Landscape of Dihydrofolate Reductase Catalysis." Science **313**(5793): 1638-1642.

Bonneau, R., J. Tsai, et al. (2001). "Rosetta in CASP4: Progress in ab initio protein structure prediction." Proteins: Structure, Function, and Bioinformatics **45**(S5): 119-126.

Bouvignies, G., P. Vallurupalli, et al. (2011). "Solution structure of a minor and transiently formed state of a T4 lysozyme mutant." Nature **477**(7362): 111-114.

Bradley, P., K. M. S. Misura, et al. (2005). "Toward High-Resolution de Novo Structure Prediction for Small Proteins." Science **309**(5742): 1868-1871.

Brady, G. P. and P. F. W. Stouten (2000). "Fast prediction and visualization of protein binding pockets with PASS." J Comput-Aided Mol Des **14**: 383 - 401.

Bray, Jenelle K., Dahlia R. Weiss, et al. (2011). "Optimized Torsion-Angle Normal Modes Reproduce Conformational Changes More Accurately Than Cartesian Modes." Biophysical Journal **101**(12): 2966-2969.

Brocklehurst, K., S. J. Willenbrock, et al. (1983). "Effects of conformational selectivity and of overlapping kinetically influential ionizations on the characteristics of pH-dependent enzyme kinetics. Implications of free-enzyme pKa variability in reactions of papain for its catalytic mechanism." Biochem. J. **211**(3): 701-708.

Brooks, B. and M. Karplus (1985). "Normal modes for specific motions of macromolecules: application to the hinge-bending mode of lysozyme." Proceedings of the National Academy of Sciences of the United States of America **82**(15): 4995-4999.

Brunger, A. T. (1992). "Free R value: a novel statistical quantity for assessing the accuracy of crystal structures." Nature **355**(6359): 472-475.

Burgen, A. S. V., G. C. K. Roberts, et al. (1975). "Binding of flexible ligands to macromolecules." Nature **253**(5494): 753-755.

Burley, S. K. (2000). "An overview of structural genomics." Nat Struct Mol Biol.

Bustamante, C., Y. R. Chemla, et al. (2004). "MECHANICAL PROCESSES IN BIOCHEMISTRY." Annual Review of Biochemistry **73**(1): 705-748.

Capra, J. A., R. A. Laskowski, et al. (2009). "Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure." PLoS Comput Biol **5**(12): e1000585.

Capra, J. A. and M. Singh (2007). "Predicting functionally important residues from sequence conservation." Bioinformatics **23**(15): 1875-1882.

Carugo, O. (2003). "How root-mean-square distance (r.m.s.d.) values depend on the resolution of protein structures that are compared." Journal of Applied Crystallography **36**(1): 125-128.

Carugo, O. and S. Pongor (2001). "A normalized root-mean-spuare distance for comparing protein three-dimensional structures." Protein Science **10**(7): 1470-1473.

Cavasotto, C. N., J. A. Kovacs, et al. (2005). "Representing Receptor Flexibility in Ligand Docking through Relevant Normal Modes." Journal of the American Chemical Society **127**(26): 9632-9640.

Chargaff, E., R. Lipshitz, et al. (1951). "The Composition of the Deoxyribonucleic Acid of Salmon Sperm." Journal of Biological Chemistry **192**(1): 223-230.

Chinea, G., G. Padron, et al. (1995). "The use of position-specific rotamers in model building by homology." Proteins: Structure, Function, and Bioinformatics **23**(3): 415-421.

Chiti, F., N. Taddei, et al. (1999). "Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding." Nat Struct Mol Biol **6**(11): 1005-1009.

Chothia, C., A. M. Lesk, et al. (1986). "The predicted structure of immunoglobulin D1.3 and its comparison with the crystal structure." Science **233**(4765): 755-758.

Cohen, F. E. and M. J. E. Sternberg (1980). "On the prediction of protein structure: The significance of the root-mean-square deviation." Journal of Molecular Biology **138**(2): 321-333.

Cole, C. and J. Warwicker (2002). "Side-chain conformational entropy at protein–protein interfaces." Protein Science **11**(12): 2860-2870.

Cole, R. and J. P. Loria (2002). "Evidence for Flexibility in the Function of Ribonuclease A." Biochemistry **41**(19): 6072-6081.

Collaborative (1994). "The CCP4 suite: programs for protein crystallography." Acta Crystallographica Section D **50**(5): 760-763.

Coutsias, E. A., C. Seok, et al. (2004). "Using quaternions to calculate RMSD." Journal of Computational Chemistry **25**(15): 1849-1857.

Crick, F. (1958). "On protein synthesis." Symp Soc Exp Biol. **12**: 138-163.

Crick, F. (1988). What Mad Pursuit: A Personal View of Scientific Discovery.

Cui, Q., G. Li, et al. (2004). "A Normal Mode Analysis of Structural Plasticity in the Biomolecular Motor F1-ATPase." Journal of Molecular Biology **340**(2): 345-372.

Dai, S., R. Friemann, et al. (2007). "Structural snapshots along the reaction pathway of ferredoxin-thioredoxin reductase." Nature **448**(7149): 92-96.

Das, R. and D. Baker (2008). "Macromolecular modeling with rosetta." Annu Rev Biochem **77**: 363-82.

Dauter, Z. and M. Dauter (1999). "Anomalous signal of solvent bromides used for phasing of lysozyme." Journal of Molecular Biology **289**(1): 93-101.

de Groot, B. L., S. Hayward, et al. (1998). "Domain motions in bacteriophage T4 lysozyme: A comparison between molecular dynamics and crystallographic data." Proteins Structure Function and Genetics **31**(2): 116-127.

Delbaere, L. T. J., G. D. Brayer, et al. (1979). "Comparison of the predicted model of [alpha]-lytic protease with the X-ray structure." Nature **279**(5709): 165-168.

DePristo, M. A., P. I. W. de Bakker, et al. (2004). "Heterogeneity and Inaccuracy in Protein Structures Solved by X-Ray Crystallography." Structure **12**(5): 831-838.

Desmet, J., M. D. Maeyer, et al. (1992). "The dead-end elimination theorem and its use in protein side-chain positioning." Nature **356**(6369): 539-542.

Dessailly, B. H., M. F. Lensink, et al. (2008). "LigASiteâ€"a database of biologically relevant binding sites in proteins with known apo-structures." Nucleic Acids Research **36**(suppl 1): D667-D673.

DiMaio, F., T. C. Terwilliger, et al. (2011). "Improved molecular replacement by density- and energy-guided protein structure optimization." Nature **473**(7348): 540-543.

Dobbins, S. E., V. I. Lesk, et al. (2008). "Insights into protein flexibility: The relationship between normal modes and conformational change upon proteinâ€"protein docking." Proceedings of the National Academy of Sciences **105**(30): 10390-10395.

Dodge, C., R. Schneider, et al. (1998). "The HSSP database of protein structureâ€"sequence alignments and family profiles." Nucleic Acids Research **26**(1): 313-315.

Doruker P, A. R. A. I. B. (2000). "Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: Application to ?-amylase inhibitor." Proteins: Structure, Function, and Genetics **40**(3): 512-524.

Durand P, G. T. Y.-H. S. (1994). "A new approach for determining low-frequency normal modes in macromolecules." Biopolymers **34**(6): 759-771.

Edwards, C. H. and D. E. Penney (1987). Elementary Linear Algebra.

Eisenmesser, E. Z., D. A. Bosco, et al. (2002). Enzyme Dynamics During Catalysis. Science. **295:** 1520-1523.

Eisenmesser, E. Z., O. Millet, et al. (2005). "Intrinsic dynamics of an enzyme underlies catalysis." Nature **438**(7064): 117-121.

Epstein, C. J., R. F. Goldberger, et al. (1963). "The Genetic Control of Tertiary Protein Structure: Studies With Model Systems." Cold Spring Harbor Symposia on Quantitative Biology **28**: 439-449.

Eriksson, A. E., W. A. Baase, et al. (1992). "A cavity-containing mutant of T4 lysozyme is stabilized by buried benzene." Nature **355**(6358): 371-373.

Eriksson, A. E., W. A. Baase, et al. (1992). "Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect." Science **255**(5041): 178-183.

Falzone, C. J., P. E. Wright, et al. (1994). "Dynamics of a flexible loop in dihydrofolate reductase from Escherichia coli and its implication for catalysis." Biochemistry **33**(2): 439-442.

Farrell, D., E. S. Miranda, et al. (2010). "Titration_DB: Storage and analysis of NMR-monitored protein pH titration curves." Proteins: Structure, Function, and Bioinformatics **78**(4): 843-857.

Fenimore, P. W., H. Frauenfelder, et al. (2002). "Slaving: Solvent fluctuations dominate protein dynamics and functions." Proceedings of the National Academy of Sciences of the United States of America **99**(25): 16047-16051.

Fersht, A. (1998). Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding, W. H. Freeman.

Fischer, E. (1894). "Einfluss der Configuration auf die Wirkung der Enzyme." Berichte der deutschen chemischen Gesellschaft **27**(3): 2985-2993.

Fisher, H. F., A. H. Colen, et al. (1981). "Temperature-dependent [Delta]C0p generated by a shift in equilibrium between macrostates of an enzyme." Nature **292**(5820): 271-272.

Flores, S., N. Echols, et al. (2006). "The Database of Macromolecular Motions: new features added at the decade mark." Nucleic Acids Research **34**(suppl 1): D296-D301.

Flory, P. J. (1976). "Statistical thermodynamics of random networks." Proc. R. Soc. Lond. A. **351**: 351-380.

Fox, R. O., P. A. Evans, et al. (1986). "Multiple conformations of a protein demonstrated by magnetization transfer NMR spectroscopy." Nature **320**(6058): 192-194.

Fraser, J. S., M. W. Clarkson, et al. (2009). "Hidden alternative structures of proline isomerase essential for catalysis." Nature **462**(7273): 669-673.

Fraser, J. S., H. van den Bedem, et al. (2011). "Accessing protein conformational ensembles using room-temperature X-ray crystallography." Proceedings of the National Academy of Sciences **108**(39): 16247-16252.

Frauenfelder, H., G. A. Petsko, et al. (1979). "Temperature-dependent X-ray diffraction as a probe of protein structural dynamics." Nature **280**(5723): 558-563.

Ganesh, V. K., S. K. Muthuvel, et al. (2005). "Structural Basis for Antagonism by Suramin of Heparin Binding to Vaccinia Complement Proteinâ€ ,â€¡." Biochemistry **44**(32): 10757-10765.

Gerstein, M. and W. Krebs (1998). "A database of macromolecular motions." Nucl. Acids Res. **26**(18): 4280-4290.

Gibbs, M. R., P. C. E. Moody, et al. (1990). "Crystal structure of the ASP-199.fwdarw.asparagine mutant of chloramphenicol acetyltransferase to 2.35.ANG. resolution: structural consequences of disruption of a buried salt bridge." Biochemistry **29**(51): 11261-11265.

Gilis, D. and M. Rooman (2000). "PoPMuSiC, an algorithm for predicting protein mutant stability changes. Application to prion proteins." Protein Engineering **13**(12): 849-856.

Go, M. and N. Go (1976). "Fluctuations of an alpha-helix." Biopolymers **15**(6): 1119-1127.

Guerois, R., J. E. Nielsen, et al. (2002). "Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of More Than 1000 Mutations." Journal of Molecular Biology **320**(2): 369-387.

Haber, E. and C. B. Anfinsen (1962). "Side-chain Interactions Governing the Pairing of Half-cystine Residues in Ribonuclease." Journal of Biological Chemistry **237:**: 1839-1844.

Haliloglu, T. and I. Bahar (1999). "Structure-based analysis of protein dynamics: Comparison of theoretical results for hen lysozyme with X-ray diffraction and NMR relaxation data." Proteins: Structure, Function, and Genetics **37**(4): 654-667.

Haliloglu, T., I. Bahar, et al. (1997). "Gaussian Dynamics of Folded Proteins." Physical Review Letters **79**(16): 3090.

Hammes, G. G., Y.-C. Chang, et al. (2009). "Conformational selection or induced fit: A flux description of reaction mechanism." Proceedings of the National Academy of Sciences **106**(33): 13737-13741.

Han, K. F. and D. Baker (1995). "Recurring Local Sequence Motifs in Proteins." Journal of Molecular Biology **251**(1): 176-187.

Hansen, D., P. Vallurupalli, et al. (2008). "Using relaxation dispersion NMR spectroscopy to determine structures of excited, invisible protein states." Journal of Biomolecular NMR **41**(3): 113-120.

Hayward S, H. J. C. B. (1998). "Systematic analysis of domain motions in proteins from conformational change: New results on citrate synthase and T4 lysozyme." Proteins: Structure, Function, and Genetics **30**(2): 144-154.

Heinz, D. W., W. A. Baase, et al. (1992). "Folding and function of a T4 lysozyme containing 10 consecutive alanines illustrate the redundancy of information in an amino acid sequence." Proceedings of the National Academy of Sciences **89**(9): 3751-3755.

Hendlich, M., F. Rippmann, et al. (1997). "LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins." Journal of Molecular Graphics and Modelling **15**(6): 359-363.

Henikoff, S. and J. G. Henikoff (1992). "Amino acid substitution matrices from protein blocks." Proceedings of the National Academy of Sciences **89**(22): 10915-10919.

Henrich, S., O. M. H. Salo-Ahen, et al. (2009). "Computational approaches to identifying and characterizing protein binding sites for ligand design." Journal of Molecular Recognition **23**(2): 209-219.

Henrick, K., Z. Feng, et al. (2008). "Remediation of the protein data bank archive." Nucl. Acids Res. **36**(suppl_1): D426-433.

Henrick, K. and J. M. Thornton (1998). "PQS: a protein quaternary structure file server." Trends in Biochemical Sciences **23**(9): 358-361.

Henzler-Wildman, K. and D. Kern (2007). "Dynamic personalities of proteins." Nature **450**(7172): 964-972.

Henzler-Wildman, K. A., M. Lei, et al. (2007). "A hierarchy of timescales in protein dynamics is linked to enzyme catalysis." Nature **450**(7171): 913-916.

Henzler-Wildman, K. A., V. Thai, et al. (2007). "Intrinsic motions along an enzymatic reaction trajectory." Nature **450**(7171): 838-844.

Hill, A. V. (1910). "The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves." The Journal of Physiology **40**(Suppl): iv-vii.

Hinsen, K. (1998). "Analysis of domain motions by approximate normal mode calculations." Proteins: Structure, Function, and Genetics **33**(3): 417-429.

Hinsen, K. A. T. M. J. F. (1999). "Analysis of domain motions in large proteins." Proteins: Structure, Function, and Genetics **34**(3): 369-382.

Huang, B. and M. Schroder (2006). "LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation." BMC Struct Biol **6**: 19 - 29.

Humphrey, W., A. Dalke, et al. (1996). "VMD: Visual molecular dynamics." Journal of Molecular Graphics **14**(1): 33-38.

J. M. Hodsdon, G. M. B., L. C. Sieker, L. H. Jensen (1990). "Refinement of triclinic lysozyme: I. Fourier and least-squares methods." Acta Crystallographica Section B **46**: 54-62.

Jackson, C. J., J. L. Foo, et al. (2009). "Conformational sampling, catalysis, and evolution of the bacterial phosphotriesterase." Proceedings of the National Academy of Sciences **106**(51): 21631-21636.

James, H. M. (1947). "Statistical Properties of Networks of Flexible Chains." The Journal of Chemical Physics **15**(9): 651-668.

Jelsch, C., M. M. Teeter, et al. (2000). "Accurate protein crystallography at ultra-high resolution: Valence electron distribution in crambin." Proceedings of the National Academy of Sciences **97**(7): 3171-3176.

Jeong, J. I., Y. Jang, et al. (2006). "A connection rule for -carbon coarse-grained elastic network models using chemical bond information." Journal of Molecular Graphics and Modelling **24**(4): 296-306.

Johnston, M. A., C. R. Søndergaard, et al. (2011). "Integrated prediction of the effect of mutations on multiple protein characteristics." Proteins: Structure, Function, and Bioinformatics **79**(1): 165-178.

Jones, J. E. (1924). "On the Determination of Molecular Fields. II. From the Equation of State of a Gas." Proceedings of the Royal Society of London. Series A **106**(738): 463-477.

Jones, T. A. and G. J. Kleywegt (2007). "Experimental Data for Structure Papers." Science **317**(5835): 194-195.

Jones, T. A., J. Y. Zou, et al. (1991). "Improved methods for building protein models in electron density maps and the location of errors in these models." Acta Crystallographica Section A **47**(2): 110-119.

Joosten, R. P., T. Womack, et al. (2009). "Re-refinement from deposited X-ray data can deliver improved models for most PDB entries." Acta Crystallographica Section D **65**(2): 176-185.

Kadirvelraj, R., N. C. Sennett, et al. (2011). "Role of Packing Defects in the Evolution of Allostery and Induced Fit in Human UDP-Glucose Dehydrogenase." Biochemistry **50**(25): 5780-5789.

Kahraman, A., R. J. Morris, et al. (2010). "On the diversity of physicochemical environments experienced by identical ligands in binding pockets of unrelated proteins." Proteins: Structure, Function, and Bioinformatics **78**(5): 1120-1136.

Kantardjieff, K. A. and B. Rupp (2003). "Matthews coefficient probabilities: Improved estimates for unit cell contents of proteins, DNA, and protein-nucleic acid complex crystals." Protein Science **12**(9): 1865-1871.

Karplus, M. and J. A. McCammon (2002). "Molecular dynamics simulations of biomolecules." Nat Struct Mol Biol **9**(9): 646-652.

Keeler, J. (2006). Understanding NMR Spectroscopy, Wiley.

Kelly, J. A., A. R. Sielecki, et al. (1979). "X-ray crystallography of the binding of the bacterial cell wall trisaccharide NAM-NAG-NAM to lysozyme." Nature **282**(5741): 875-878.

Kempf, J. G. and J. P. Loria (2004). Measurement of Intermediate Exchange Phenomena. **278:** 185-231.

Kitao, A. and N. Go (1999). "Investigating protein dynamics in collective coordinate space." Current Opinion in Structural Biology **9**(2): 164-169.

Kleywegt, G. J. (1996). "Use of Non-crystallographic Symmetry in Protein Structure Refinement." Acta Crystallographica Section D **52**(4): 842-857.

Kleywegt, G. J. and T. A. Jones (1996). "Efficient Rebuilding of Protein Structures." Acta Crystallographica Section D **52**(4): 829-832.

Kleywegt, G. J. and T. A. Jones (1996). "Phi/Psi-chology: Ramachandran revisited." Structure **4**(12): 1395-1400.

Koch, C., A. Heine, et al. (2011). "Ligand-induced fit affects binding modes and provokes changes in crystal packing of aldose reductase." Biochimica et Biophysica Acta (BBA) - General Subjects **1810**(9): 879-887.

Kondrashov, D. A., A. W. Van Wynsberghe, et al. (2007). "Protein Structural Variation in Computational Models and Crystallographic Data." Structure (London, England : 1993) **15**(2): 169-177.

Korkegian, A., M. E. Black, et al. (2005). "Computational Thermostabilization of an Enzyme." Science **308**(5723): 857-860.

Kortemme, T. and D. Baker (2002). "A simple physical model for binding energy hot spots in proteinâ€"protein complexes." Proceedings of the National Academy of Sciences **99**(22): 14116-14121.

Korzhnev, D. M., T. L. Religa, et al. (2010). "A Transient and Low-Populated Protein-Folding Intermediate at Atomic Resolution." Science **329**(5997): 1312-1316.

Koshland, D. E. (1958). "Application of a Theory of Enzyme Specificity to Protein Synthesis." Proceedings of the National Academy of Sciences **44**(2): 98-104.

Kosloff, M. and R. Kolodny (2008). "Sequence-similar, structure-dissimilar protein pairs in the PDB." Proteins: Structure, Function, and Bioinformatics **71**(2): 891-902.

Kovrigin, E. L. and J. P. Loria (2006). "Enzyme Dynamics along the Reaction Coordinate:â€‰ Critical Role of a Conserved Residue." Biochemistry **45**(8): 2636-2647.

Krebs, W. G., V. Alexandrov, et al. (2002). "Normal mode analysis of macromolecular motions in a database framework: Developing mode concentration as a useful classifying statistic." Proteins: Structure, Function, and Genetics **48**(4): 682-695.

Krieger, E., K. Joo, et al. (2009). "Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in CASP8." Proteins **77 Suppl 9**: 114-22.

Krissinel, E. and K. Henrick (2007). "Inference of Macromolecular Assemblies from Crystalline State." Journal of Molecular Biology **372**(3): 774-797.

Krivov, G. G., M. V. Shapovalov, et al. (2009). "Improved prediction of protein side-chain conformations with SCWRL4." Proteins: Structure, Function, and Bioinformatics **77**(4): 778-795.

Kukic, P., D. Farrell, et al. (2009). "Improving the analysis of NMR spectra tracking pH-induced conformational changes: Removing artefacts of the electric field on the NMR chemical shift." Proteins: Structure, Function, and Bioinformatics **78**(4): 971-984.

Ladurner, A. G. and A. R. Fersht (1999). "Upper limit of the time scale for diffusion and chain collapse in chymotrypsin inhibitor 2." Nat Struct Mol Biol **6**(1): 28-31.

Lang, P. T., H. L. Ng, et al. (2010). "Automated electron-density sampling reveals widespread conformational polymorphism in proteins." Protein Science **19**(7): 1420-1431.

Lange, O. F., N.-A. Lakomek, et al. (2008). "Recognition Dynamics Up to Microseconds Revealed from an RDC-Derived Ubiquitin Ensemble in Solution." Science **320**(5882): 1471-1475.

LaPlante, S. R., J. R. Gillard, et al. (2010). "Importance of Ligand Bioactive Conformation in the Discovery of Potent Indole-Diamide Inhibitors of the Hepatitis C Virus NS5B." Journal of the American Chemical Society **132**(43): 15204-15212.

Laskowski, R. A. (1995). "SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions." J Mol Graph **13**: 323 - 330.

Laskowski, R. L., NM; Swindells, MB; Thornton, JM (1996). "Protein clefts in molecular recognition and function." Protein Science **5**: 2438-2452.

Lazaridis, T. and M. Karplus (1999). "Effective energy function for proteins in solution." Proteins: Structure, Function, and Bioinformatics **35**(2): 133-152.

Levitt, D. G. and L. J. Banaszak (1992). "POCKET: A computer graphics method for identifying and displaying protein cavities and their surrounding amino acids." Journal of Molecular Graphics **10**(4): 229-234.

Levitt, M., C. Sander, et al. (1985). "Protein normal-mode dynamics: Trypsin inhibitor, crambin, ribonuclease and lysozyme." Journal of Molecular Biology **181**(3): 423-447.

Li, G. and Q. Cui (2002). "A Coarse-Grained Normal Mode Approach for Macromolecules: An Efficient Implementation and Application to Ca2+-ATPase." Biophys. J. **83**(5): 2457-2474.

Li, G. and Q. Cui (2004). "Analysis of Functional Motions in Brownian Molecular Machines with an Efficient Block Normal Mode Approach: Myosin-II and Ca2+-ATPase." Biophys. J. **86**(2): 743-763.

Li, W. and A. Godzik (2006). "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences." Bioinformatics **22**(13): 1658-1659.

Liang, J., C. Woodward, et al. (1998). "Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design." Protein Science **7**(9): 1884-1897.

Lifson, S. and A. Warshel (1968). "Consistent Force Field for Calculations of Conformations, Vibrational Spectra, and Enthalpies of Cycloalkane and n-Alkane Molecules." The Journal of Chemical Physics **49**(11): 5116-5129.

Lim, K., A. Nadarajah, et al. (1998). "Locations of Bromide Ions in Tetragonal Lysozyme Crystals." Acta Crystallographica Section D **54**(5): 899-904.

Lin, J. (1991). "Divergence measures based on the shannon entropy." IEEE Transactions on Information Theory **37**: 145-151.

Lindahl, E. and M. Delarue (2005). "Refinement of docked proteinâ€“ligand and proteinâ€“DNA structures using low frequency normal mode amplitude optimization." Nucleic Acids Research **33**(14): 4496-4506.

Lindorff-Larsen, K., R. B. Best, et al. (2005). "Simultaneous determination of protein structure and dynamics." Nature **433**: 128-132.

Lindorff-Larsen, K. and J. Ferkinghoff-Borg (2009). "Similarity Measures for Protein Ensembles." PLoS ONE **4**(1): e4203.

Lindorff-Larsen, K., S. Piana, et al. (2011). "How Fast-Folding Proteins Fold." Science **334**(6055): 517-520.

Liu, L., M. L. Quillin, et al. (2008). "Use of experimental crystallographic phases to examine the hydration of polar and nonpolar cavities in T4 lysozyme." Proceedings of the National Academy of Sciences **105**(38): 14406-14411.

Liu, Y. and D. Eisenberg (2002). "3D domain swapping: As domains continue to swap." Protein Sci **11**(6): 1285-1299.

Lu, M. and J. Ma (2005). "The Role of Shape in Determining Molecular Motions." **89**(4): 2395-2401.

Lumry, R. and H. Eyring (1954). "Conformation Changes of Proteins." The Journal of Physical Chemistry **58**(2): 110-120.

M. Jack Borrok, L. L. K. K. T. F. (2007). "Conformational changes of glucose/galactose-binding protein illuminated by open, unliganded, and ultra-high-resolution ligand-bound structures." Protein Science **16**(6): 1032-1041.

MacCallum, J. L., L. Hua, et al. (2009). "Assessment of the protein-structure refinement category in CASP8." Proteins: Structure, Function, and Bioinformatics **77**(S9): 66-80.

Mackereth, C. D., T. Madl, et al. (2011). "Multi-domain conformational selection underlies pre-mRNA splicing regulation by U2AF." Nature **475**(7356): 408-411.

Maiorov, V. N. and G. M. Crippen (1995). "Size-independent comparison of protein three-dimensional structures." Proteins: Structure, Function, and Genetics **22**(3): 273-283.

Marques, O. and Y.-H. Sanejouand (1995). "Hinge-bending motion in citrate synthase arising from normal mode calculations." Proteins: Structure, Function, and Genetics **23**(4): 557-560.

Masterson, L. R., C. Cheng, et al. (2010). "Dynamics connect substrate recognition to catalysis in protein kinase A." Nat Chem Biol **6**(11): 821-828.

Matthews, B. W. (1968). "Solvent content of protein crystals." Journal of Molecular Biology **33**(2): 491-497.

Matthews, B. W., H. Nicholson, et al. (1987). "Enhanced protein thermostability from site-directed mutations that decrease the entropy of unfolding." Proceedings of the National Academy of Sciences **84**(19): 6663-6667.

May, A. and M. Zacharias (2008). "Proteinâˆ'Ligand Docking Accounting for Receptor Side Chain and Global Flexibility in Normal Modes: Evaluation on Kinase Inhibitor Cross Docking." Journal of Medicinal Chemistry **51**(12): 3499-3506.

McCammon, J. A., B. R. Gelin, et al. (1977). "Dynamics of folded proteins." Nature **267**(5612): 585-590.

McCammon, J. A., B. R. Gelin, et al. (1976). "The hinge-bending mode in lysozyme." Nature **262**(5566): 325-326.

McCammon, J. A. and S. H. Northrup (1981). "Gated binding of ligands to proteins." Nature **293**(5830): 316-317.

McElheny, D., J. R. Schnell, et al. (2005). "Defining the role of active-site loop fluctuations in dihydrofolate reductase catalysis." Proceedings of the National Academy of Sciences of the United States of America **102**(14): 5032-5037.

Meireles, L., M. Gur, et al. (2011). "Pre-existing soft modes of motion uniquely defined by native contact topology facilitate ligand binding to proteins." Protein Science **20**(10): 1645-1658.

Milburn, M. V., G. G. Prive, et al. (1991). "Three-dimensional structures of the ligand-binding domain of the bacterial aspartate receptor with and without a ligand." Science **254**(5036): 1342-1347.

Milburn, M. V., L. Tong, et al. (1990). "Molecular switch for signal transduction: structural differences between active and inactive forms of protooncogenic ras proteins." Science **247**(4945): 939-945.

Miller, H., S. S. Mande, et al. (1995). "An L40C Mutation Converts the Cysteine-Sulfenic Acid Redox Center in Enterococcal NADH Peroxidase to a Disulfide." Biochemistry **34**(15): 5180-5190.

Mitra, K., C. Schaffitzel, et al. (2005). "Structure of the E. coli protein-conducting channel bound to a translating ribosome." Nature **438**(7066): 318-324.

Miyashita, O., C. Gorba, et al. (2011). "Structure modeling from small angle X-ray scattering data with elastic network normal mode analysis." Journal of Structural Biology **173**(3): 451-460.

Monod, J., J. Wyman, et al. (1965). "On the nature of allosteric transitions: A plausible model." Journal of Molecular Biology **12**(1): 88-118.

Morton, A. and B. W. Matthews (1995). "Specificity of ligand binding in a buried nonpolar cavity of T4 lysozyme: Linkage of dynamics and structural plasticity." Biochemistry **34**(27): 8576-8588.

Mowbray, S. L., C. Helgstrand, et al. (1999). "Errors and reproducibility in electron-density map interpretation." Acta Crystallographica Section D **55**(7): 1309-1319.

Mozzarelli, A., C. Rivetti, et al. (1991). "Crystals of haemoglobin with the T quaternary structure bind oxygen noncooperatively with no Bohr effect." Nature **351**(6325): 416-419.

Mulder, F. A. A., B. Hon, et al. (2002). "Slow internal dynamics in proteins: application of NMR relaxation dispersion spectroscopy to methyl groups in a cavity mutant of T4 lysozyme." J. Am. Chem. Soc **124**(7): 1443–1451.

Mulder, F. A. A., A. Mittermaier, et al. (2001). "Studying excited states of proteins by NMR spectroscopy." Nat Struct Mol Biol **8**(11): 932-935.

Murshudov, G. N., A. A. Vagin, et al. (1997). "Refinement of Macromolecular Structures by the Maximum-Likelihood Method." Acta Crystallographica Section D **53**(3): 240-255.

Murthy, K. H. M., S. A. Smith, et al. (2001). "Crystal Structure of a Complement Control Protein that Regulates Both Pathways of Complement Activation and Binds Heparan Sulfate Proteoglycans." Cell **104**(2): 301-311.

Murzin, A. G. (1999). "Structure classification-based assessment of CASP3 predictions for the fold recognition targets." Proteins: Structure, Function, and Bioinformatics **37**(S3): 88-103.

Murzin, A. G., S. E. Brenner, et al. (1995). "SCOP: A structural classification of proteins database for the investigation of sequences and structures." Journal of Molecular Biology **247**(4): 536-540.

Nagel, Z. D. and J. P. Klinman (2009). "A 21st century revisionist's view at a turning point in enzymology." Nat Chem Biol **5**(8): 543-550.

Neuvirth, H., R. Raz, et al. (2004). "ProMate: A Structure Based Prediction Program to Identify the Location of Proteinâ€“Protein Binding Sites." Journal of Molecular Biology **338**(1): 181-199.

Nicholson, H., W. J. Becktel, et al. (1988). "Enhanced protein thermostability from designed mutations that interact with [alpha]-helix dipoles." Nature **336**(6200): 651-656.

Niu, X., L. Bruschweiler-Li, et al. (2011). "Arginine Kinase: Joint Crystallographic and NMR RDC Analyses Link Substrate-Associated Motions to Intrinsic Flexibility." Journal of Molecular Biology **405**(2): 479-496.

Orengo, C. A., N. P. Brown, et al. (1992). "Fast structure alignment for protein databank searching." Proteins: Structure, Function, and Bioinformatics **14**(2): 139-167.

Orengo, C. A., A. D. Michie, et al. (1997). "CATH a hierarchic classification of protein domain structures." Structure (London, England : 1993) **5**(8): 1093-1109.

Ortiz, A. R., C. E. M. Strauss, et al. (2002). "MAMMOTH (Matching molecular models obtained from theory): An automated method for model comparison." Protein Science **11**(11): 2606-2621.

Ostermann, A., R. Waschipky, et al. (2000). "Ligand binding and conformational motions in myoglobin." Nature **404**(6774): 205-208.

Padlan, E. A., E. W. Silverton, et al. (1989). "Structure of an antibody-antigen complex: crystal structure of the HyHEL-10 Fab-lysozyme complex." Proceedings of the National Academy of Sciences **86**(15): 5938-5942.

Palmer III, A. G., C. D. Kroenke, et al. (2001). [10] Nuclear magnetic resonance methods for quantifying microsecond-to-millisecond motions in biological macromolecules. Methods in Enzymology, Academic Press. **Volume 339:** 204-238.

Perutz, M. F., M. G. Rossmann, et al. (1960). "Structure of Haemoglobin: A Three-Dimensional Fourier Synthesis at 5.5-[angst]. Resolution, Obtained by X-Ray Analysis." Nature **185**(4711): 416-422.

Petty, T. J., S. Emamzadah, et al. (2011). "An induced fit mechanism regulates p53 DNA binding kinetics to confer sequence specificity." EMBO J **30**(11): 2167-2176.

Pisliakov, A. V., J. Cao, et al. (2009). "Enzyme millisecond conformational dynamics do not catalyze the chemical step." Proceedings of the National Academy of Sciences **106**(41): 17359-17364.

Piszkiewicz, D. (1974). "pH-dependent conformational change of gastrin." Nature **248**(5446): 341-342.

Pjura, P. E., M. Matsumura, et al. (1990). "Structure of a thermostable disulfide-bridge mutant of phage T4 lysozyme shows that an engineered cross-link in a flexible region does not increase the rigidity of the folded protein." Biochemistry **29**(10): 2592-2598.

Polgár, L. and P. HalÁSz (1978). "Evidence for Multiple Reactive Forms of Papain." European Journal of Biochemistry **88**(2): 513-521.

Popovych, N., S. Sun, et al. (2006). "Dynamically driven protein allostery." Nat Struct Mol Biol **13**(9): 831-838.

Porter, C. T., G. J. Bartlett, et al. (2004). "The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data." Nucleic Acids Research **32**(suppl 1): D129-D133.

Qian, B., S. Raman, et al. (2007). "High-resolution structure prediction and the crystallographic phase problem." Nature **450**(7167): 259-264.

Rajagopalan, P. T. R., S. Lutz, et al. (2002). "Coupling Interactions of Distal Residues Enhance Dihydrofolate Reductase Catalysis: Mutational Effects on Hydride Transfer Rates." Biochemistry **41**(42): 12618-12628.

Randy, J. R. and C. Gayatri (2007). "Assessment of CASP7 predictions in the high accuracy template-based modeling category." Proteins: Structure, Function, and Bioinformatics **69**(S8): 27-37.

Rasmussen, B. F., A. M. Stock, et al. (1992). "Crystalline ribonuclease A loses function below the dynamical transition at 220 K." Nature **357**(6377): 423-424.

Read, Randy J., Paul D. Adams, et al. (2011). "A New Generation of Crystallographic Validation Tools for the Protein Data Bank." Structure (London, England : 1993) **19**(10): 1395-1412.

Reva, B. A., A. V. Finkelstein, et al. (1998). "What is the probability of a chance prediction of a protein structure with an rmsd of 6 å?" Folding and Design **3**(2): 141-147.

Rini, J. M., U. Schulze-Gahmen, et al. (1992). "Structural evidence for induced fit as a mechanism for antibody-antigen recognition." Science **255**(5047): 959-965.

Roche, S. p., S. p. Bressanelli, et al. (2006). "Crystal Structure of the Low-pH Form of the Vesicular Stomatitis Virus Glycoprotein G." Science **313**(5784): 187-191.

Ross, S. A., C. A. Sarisky, et al. (2001). "Designed protein G core variants fold to native-like structures: Sequence selection by ORBIT tolerates variation in backbone specification." Protein Science **10**(2): 450-454.

Rossmann, M. (1967). "Application of the molecular replacement equations to the heavy atom technique." Acta Crystallographica **23**(1): 173-174.

Rossmann, M. G. (2001). "Molecular replacement - historical background." Acta Crystallographica Section D **57**(10): 1360-1366.

Rupp, B. (2009). Biomolecular Crystallography, Garland Science.

Sakurai, K. and Y. Goto (2007). "Principal component analysis of the pH-dependent conformational transitions of bovine Î²-lactoglobulin monitored by heteronuclear NMR." Proceedings of the National Academy of Sciences **104**(39): 15346-15351.

Schmidt, A., M. Teeter, et al. (2011). "Crystal structure of small protein crambin at 0.48 A resolution." Acta Crystallographica Section F **67**(4): 424-428.

Schreiber, G. and A. R. Fersht (1995). "Energetics of protein-protein interactions: Analysis ofthe Barnase-Barstar interface by single mutations and double mutant cycles." Journal of Molecular Biology **248**(2): 478-486.

Schueler-Furman, O., C. Wang, et al. (2005). "Progress in Modeling of Protein Structures and Interactions." Science **310**(5748): 638-642.

Schulz, G. E., C. D. Barry, et al. (1974). "Comparison of predicted and experimentally determined secondary structure of adenyl kinase." Nature **250**(5462): 140-142.

Schwartz, S. D. and V. L. Schramm (2009). "Enzymatic transition states and dynamic motion in barrier crossing." Nat Chem Biol **5**(8): 551-558.

Seeliger, D. and B. L. de Groot (2010). "Conformational Transitions upon Ligand Binding: Holo-Structure Prediction from Apo Conformations." PLoS Comput Biol **6**(1): e1000634.

Selzer, T., S. Albeck, et al. (2000). "Rational design of faster associating and tighter binding protein complexes." Nat Struct Mol Biol **7**(7): 537-541.

Shaw, D. E., P. Maragakis, et al. (2010). "Atomic-Level Characterization of the Structural Dynamics of Proteins." Science **330**(6002): 341-346.

Sheffler, W. and D. Baker (2009). "RosettaHoles: Rapid assessment of protein core packing for structure prediction, refinement, design, and validation." Protein Science **18**(1): 229-239.

Shen, Y. and A. Bax (2007). "Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology." Journal of Biomolecular NMR **38**(4): 289-302.

Shen, Y., O. Lange, et al. (2008). "Consistent blind protein structure generation from NMR chemical shift data." Proceedings of the National Academy of Sciences **105**(12): 4685-4690.

Shindyalov, I. N. and P. E. Bourne (1998). "Protein structure alignment by incremental combinatorial extension (CE) of the optimal path." Protein Engineering **11**(9): 739-747.

Silva, D.-A., G. R. Bowman, et al. (2011). "A Role for Both Conformational Selection and Induced Fit in Ligand Binding by the LAO Protein." PLoS Comput Biol **7**(5): e1002054.

Simon, C. L., J. M. Word, et al. (2000). "The penultimate rotamer library." Proteins: Structure, Function, and Bioinformatics **40**(3): 389-408.

Smith, C. A. and T. Kortemme (2008). "Backrub-Like Backbone Simulation Recapitulates Natural Protein Conformational Variability and Improves Mutant Side-Chain Prediction." Journal of Molecular Biology **380**(4): 742-756.

Sprang, S., T. Standing, et al. (1987). "The three-dimensional structure of Asn102 mutant of trypsin: role of Asp102 in serine protease catalysis." Science **237**(4817): 905-909.

Stanfield, R. L., T. M. Fieser, et al. (1990). "Crystal structures of an antibody to a peptide and its complex with peptide antigen at 2.8 A." Science **248**(4956): 712-719.

Stock, A. M., J. M. Mottonen, et al. (1989). "Three-dimensional structure of CheY, the response regulator of bacterial chemotaxis." Nature **337**(6209): 745-749.

Suhre, K. and Y.-H. Sanejouand (2004). "On the potential of normal-mode analysis for solving difficult molecular-replacement problems." Acta Crystallographica Section D **60**(4): 796-799.

Takayama, Y. and M. Nakasako (2011). "A few low-frequency normal modes predominantly contribute to conformational responses of hen egg white lysozyme in the tetragonal crystal to variations of molecular packing controlled by environmental humidity." Biophysical Chemistry **In Press, Corrected Proof**.

Taly, A., P.-J. Corringer, et al. (2006). "Implications of the quaternary twist allosteric model for the physiology and pathology of nicotinic acetylcholine receptors." Proceedings of the National Academy of Sciences **103**(45): 16965-16970.

Tama, F. and C. L. Brooks Iii (2002). "The Mechanism and Pathway of pH Induced Swelling in Cowpea Chlorotic Mottle Virus." Journal of Molecular Biology **318**(3): 733-747.

Tama, F., M. Feig, et al. (2005). "The Requirement for Mechanical Coupling Between Head and S2 Domains in Smooth Muscle Myosin ATPase Regulation and its Implications for Dimeric Motor Function." Journal of Molecular Biology **345**(4): 837-854.

Tama, F., O. Miyashita, et al. (2004). "Flexible Multi-scale Fitting of Atomic Structures into Low-resolution Electron Density Maps with Elastic Network Normal Mode Analysis." Journal of Molecular Biology **337**(4): 985-999.

Tama, F., O. Miyashita, et al. (2004). "Normal mode based flexible fitting of high-resolution structure into low-resolution experimental data from cryo-EM." Journal of Structural Biology **147**(3): 315-326.

Tama, F. and Y. H. Sanejouand (2001). "Conformational change of proteins arising from normal mode calculations." Protein Eng. **14**(1): 1-6.

Tama, F., M. Valle, et al. (2003). "Dynamic reorganization of the functionally active ribosome explored by normal mode analysis and cryo-electron microscopy." Proceedings of the National Academy of Sciences **100**(16): 9319-9323.

Tama, F., W. Wriggers, et al. (2002). "Exploring Global Distortions of Biological Macromolecules and Assemblies from Low-resolution Structural Information and Elastic Network Theory." Journal of Molecular Biology **321**(2): 297-305.

Teeter, M. M., A. Yamano, et al. (2001). "On the nature of a glassy state of matter in a hydrated protein: Relation to protein function." Proceedings of the National Academy of Sciences **98**(20): 11242-11247.

Teilum, K., F. M. Poulsen, et al. (2006). "The inverted chevron plot measured by NMR relaxation reveals a native-like unfolding intermediate in acyl-CoA binding protein." Proceedings of the National Academy of Sciences **103**(18): 6877-6882.

Tews, I., F. Findeisen, et al. (2005). "The Structure of a pH-Sensing Mycobacterial Adenylyl Cyclase Holoenzyme." Science **308**(5724): 1020-1023.

The UniProt, C. (2011). "Ongoing and future developments at the Universal Protein Resource." Nucleic Acids Research **39**(suppl 1): D214-D219.

Tirion, M. M. (1996). "Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis." Physical Review Letters **77**(9): 1905.

Todd, A. E., R. L. Marsden, et al. (2005). "Progress of Structural Genomics Initiatives: An Analysis of Solved Target Structures." Journal of Molecular Biology **348**(5): 1235-1260.

Tokuriki, N. and D. S. Tawfik (2009). "Protein Dynamism and Evolvability." Science **324**(5924): 203-207.

Tronrud, D. E. and B. W. Matthews (2009). "Sorting the chaff from the wheat at the PDB." Protein Science **18**(1): 2-5.

Tynan-Connolly, B. M. and J. E. Nielsen (2006). "pKD: re-designing protein pKa values." Nucl. Acids Res. **34**(suppl_2): W48-51.

Tynan-Connolly, B. M. and J. E. Nielsen (2007). "Redesigning protein pKa values." Protein Science **16**(2): 239-249.

Ulrich, E. L., H. Akutsu, et al. (2008). "BioMagResBank." Nucl. Acids Res. **36**(suppl_1): D402-408.

Vagin, A. A., R. A. Steiner, et al. (2004). "REFMAC5 dictionary: organization of prior chemical knowledge and guidelines for its use." Acta Crystallographica Section D **60**(12 Part 1): 2184-2195.

Vaney, M. C., I. Broutin, et al. (2001). "Structural effects of monovalent anions on polymorphic lysozyme crystals." Acta Crystallographica Section D **57**(7): 929-940.

Veltman, O. R., V. G. H. Eijsink, et al. (1998). "Probing Catalytic Hinge Bending Motions in Thermolysin-like Proteases by Glycine â†’ Alanine Mutations." Biochemistry **37**(15): 5305-5311.

Vértessy, B. G. and F. Orosz (2010). "From “fluctuation fit” to “conformational selection”: Evolution, rediscovery, and integration of a concept." BioEssays **33**(1): 30-34.

Vitkup, D., D. Ringe, et al. (2000). "Solvent mobility and the protein 'glass' transition." Nat Struct Mol Biol **7**(1): 34-38.

Vocadlo, D. J., G. J. Davies, et al. (2001). "Catalysis by hen egg-white lysozyme proceeds via a covalent intermediate." Nature **412**(6849): 835-838.

Volkman, B. F., D. Lipson, et al. (2001). "Two-State Allosteric Behavior in a Single-Domain Signaling Protein." Science **291**(5512): 2429-2433.

Vriend, G. and C. Sander (1991). "Detection of common three-dimensional substructures in proteins." Proteins: Structure, Function, and Bioinformatics **11**(1): 52-58.

Vyas, N. K., M. N. Vyas, et al. (1988). "Sugar and signal-transducer binding sites of the Escherichia coli galactose chemoreceptor protein." Science **242**(4883): 1290-1295.

Wako, H. and S. Endo (2011). "Ligand-induced conformational change of a protein reproduced by a linear combination of displacement vectors obtained from normal mode analysis." Biophysical Chemistry **159**(2-3): 257-266.

Wako, H. and S. Endo (2011). "Ligand-induced conformational change of a protein reproduced by a linear combination of displacement vectors obtained from normal mode analysis." Biophysical Chemistry **In Press, Corrected Proof**.

Wang, H., S. Chumnarnsilpa, et al. (2009). "Helix Straightening as an Activation Mechanism in the Gelsolin Superfamily of Actin Regulatory Proteins." Journal of Biological Chemistry **284**(32): 21265-21269.

Wang, L., N. M. Goodey, et al. (2006). "Coordinated effects of distal mutations on environmentally coupled tunneling in dihydrofolate reductase." Proceedings of the National Academy of Sciences **103**(43): 15753-15758.

Wang, M., R. T. Borchardt, et al. (2005). "Domain Motions and the Open-to-Closed Conformational Transition of an Enzyme:â€‰ A Normal Mode Analysis of S-Adenosyl-l-homocysteine Hydrolaseâ€ " Biochemistry **44**(19): 7228-7239.

Weaver, L. H. and B. W. Matthews (1987). "Structure of bacteriophage T4 lysozyme refined at 1.7 Å resolution." Journal of Molecular Biology **193**(1): 189-199.

Weinert, E. E., C. M. Phillips-Piro, et al. (2011). "Controlling Conformational Flexibility of an O2-Binding H-NOX Domain." Biochemistry: null-null.

Weisel, M., E. Proschak, et al. (2007). "PocketPicker: analysis of ligand binding-sites with shape descriptors." Chemistry Central Journal **1**(1): 7.

Weiss, M. (2001). "Global indicators of X-ray data quality." Journal of Applied Crystallography **34**(2): 130-135.

Wells, S., S. Menor, et al. (2005). "Constrained geometric simulation of diffusive motion in proteins." Phys. Biol **2**: 1-10.

Williams, J. C. and A. E. McDermott (1995). "Dynamics of the Flexible Loop of Triose-Phosphate Isomerase: The Loop Motion Is Not Ligand Gated." Biochemistry **34**(26): 8309-8319.

Wilson, M. A. and A. T. Brunger (2000). "The 1.0 Å crystal structure of Ca2+-bound calmodulin: an analysis of disorder and implications for functionally relevant plasticity." Journal of Molecular Biology **301**(5): 1237-1256.

Winn, M. D., C. C. Ballard, et al. "Overview of the CCP4 suite and current developments." Acta Crystallographica Section D **67**(4): 235-242.

Winn, M. D., M. N. Isupov, et al. (2001). "Use of TLS parameters to model anisotropic displacements in macromolecular refinement." Acta Crystallographica Section D **57**(1): 122-133.

Wolf-Watz, M., V. Thai, et al. (2004). "Linkage between dynamics and catalysis in a thermophilic-mesophilic enzyme pair." Nat Struct Mol Biol **11**(10): 945-949.

Wrabl, J. O., J. Gu, et al. (2011). "The role of protein conformational fluctuations in allostery, function, and evolution." Biophysical Chemistry **In Press, Corrected Proof**.

Wray, J. W., W. A. Baase, et al. (1999). "Structural analysis of a non-contiguous second-site revertant in T4 lysozyme shows that increasing the rigidity of a protein can enhance its stability." Journal of Molecular Biology **292**(5): 1111-1120.

Yang, L.-W. and I. Bahar (2005). "Coupling between Catalytic Site and Collective Dynamics: A Requirement for Mechanochemical Activity of Enzymes." Structure **13**(6): 893-904.

Yang, S., L. Blachowicz, et al. (2010). "Multidomain assembled states of Hck tyrosine kinase in solution." Proceedings of the National Academy of Sciences **107**(36): 15757-15762.

Yankeelov, J. A. and D. E. Koshland (1965). "Evidence for Conformation Changes Induced by Substrates of Phosphoglucomutase." Journal of Biological Chemistry **240**(4): 1593-1602.

Ye, Y. and A. Godzik (2004). "FATCAT: a web server for flexible structure comparison and structure similarity searching." Nucleic Acids Research **32**(suppl 2): W582-W585.

Yu, J., Y. Zhou, et al. (2010). "Roll: a new algorithm for the detection of protein pockets and cavities with a rolling probe sphere." Bioinformatics **26**(1): 46-52.

Zacharias, M. and H. Sklenar (1999). "Harmonic modes as variables to approximately account for receptor flexibility in ligand–receptor docking simulations: Application to DNA minor groove ligand complex." Journal of Computational Chemistry **20**(3): 287-300.

Zand, R., B. B. L. Agrawal, et al. (1971). "pH-Dependent Conformational Changes of Concanavalin A." Proceedings of the National Academy of Sciences **68**(9): 2173-2176.

Zar, J. H. (1998). Biostatistical Analysis, Ch. 8, Two-Sample Hypotheses, Prentice Hall.

Zar, J. H. (1998). Biostatistical Analysis, Ch. 19, Simple Linear Correlation, Prentice Hall.

Zemla, A. (2003). "LGA: a method for finding 3D similarities in protein structures." Nucl. Acids Res. **31**(13): 3370-3374.

Zhang, Y. and J. Skolnick (2005). "TM-align: a protein structure alignment algorithm based on the TM-score." Nucleic Acids Research **33**(7): 2302-2309.

Zheng, W. and B. R. Brooks (2005). "Probing the Local Dynamics of Nucleotide-Binding Pocket Coupled to the Global Dynamics: Myosin versus Kinesin." **89**(1): 167-178.

Zheng, W. and S. Doniach (2003). "A comparative study of motor-protein motions by using a simple elastic-network model." PNAS **100**(23): 13253-13258.

Zhou, H.-X., S. T. Wlodek, et al. (1998). "Conformation gating as a mechanism for enzyme specificity." Proceedings of the National Academy of Sciences **95**(16): 9280-9283.