UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Structural basis and dynamics of signal transduction in the PAS protein family

Vreede, J.

Publication date
2007

Document Version
Final published version

Link to publication

Citation for published version (APA):
Vreede, J. (2007). *Structural basis and dynamics of signal transduction in the PAS protein family*. [Thesis, fully internal, Universiteit van Amsterdam].

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:09 Mar 2023

Structural basis and dynamics of signal transduction in the PAS protein family

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. mr. P.F. van der Heijden
ten overstaan van een door het college voor promoties ingestelde
commissie, in het openbaar te verdedigen in de Aula der Universiteit

op donderdag 15 februari 2007, te 12.00 uur

door

Jocelyne Vreede
geboren te Geleen

Promotiecomissie

Promotores:	prof. dr. K.J. Hellingwerf
	prof. dr. W. Crielaard
Overige leden:	prof. dr. D.M.F. van Aalten
	prof. dr. H.J.C. Berendsen
	prof. dr. P.G. Bolhuis
	prof. dr. D. Frenkel
	dr. H.C.J. Hoefsloot
	prof. dr. C.G. de Koster
	dr. C. Oostenbrink

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

Structural basis and dynamics of signal transduction in the PAS protein family

Jocelyne Vreede

The research presented in this thesis was performed in the Swammerdam Institute for Life Sciences.

This document was created in latex.

Printed by PrintPartners Ipskamp B.V., Enschede

Cover: *Signal transduction* by Jocelyne Vreede, 2006

voor Lianne en Steven

Contents

1 Signal transduction in cells **7**
Signal transduction in micro-organisms 8
Signal transduction in higher organisms 9
Sensor characteristics 11
Outline of this thesis 12

2 Investigating the structure and dynamics of PAS domains **15**
Introduction 15
Materials and methods 16
Results and discussion 20
Concluding remarks 24

3 Predicting the signaling state of Photoactive Yellow Protein **25**
Introduction 25
Methods 28
Results 29
Discussion 35
Concluding remarks 38

4 Conformational requirements for an efficient recovery reaction of Photoactive Yellow Protein **41**
Introduction 41
Methods 42
Results and discussion 46
Concluding remarks 54

5 Protein Triads: A new method to analyse collective motions in proteins **57**
Introduction 58
Theory 59
Computational details 61
Results and discussion 64
Conclusion 71

6 Auto-inducer mediates TraR DNA binding through backbone fluctuations 73
Introduction 73
Methods 76
Results and discussion 77
Conclusion 87

7 Discussion 89

A Protein structure 95

B Protein Dynamics 105

Bibliography 112

Summary 127

Samenvatting 131

Curriculum vitae 135

Publications 137

Nawoord 139

Chapter 1

Signal transduction in cells

Cells interact with their environment, whether they are a building block of higher organisms, or free-living organisms. To successfully compete in its ecosystem, any cell has to continuously adapt to changes in external conditions. This requires the cell to perceive changes in its environment and subsequently generate an appropriate response to those changes, a process called signal transduction.

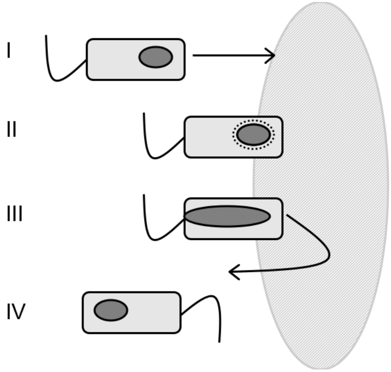


Figure 1.1: **Cartoon of a bacterium responding to its environment.** I. A bacterium swims toward a unfavorable zone. II. A protein within the bacterium perceives the inhibitory conditions and becomes activated. III. Activation leads to a change in the shape of the protein. As a consequence, it switches on other machinery in the bacterium: the flagellum. IV. The bacterium turns around and swims away from the region with the unfavorable conditions.

Signal transduction in micro-organisms

Particularly for organisms as small as bacteria it is relevant to be able to respond to their surroundings, because of their high surface to volume ratio. Bacteria live in environments that can change rapidly and unpredictably in nutrient and toxin levels, acidity, temperature, osmolarity, humidity and many other conditions. To survive, the organisms must constantly monitor the external conditions and adapt to them.

Small environmental changes are detected and trigger changes in gene expression and motility to enhance survival. Extracellullar chemical and/or physical stimuli elicit an intracellular response: expression and/or inhibition of genes (the genomic response), activation of the flagellum (the locomotor response) and activation and/or inhibition of enzymes (the biochemical response). As is fundamental to any cellular signaling system, the tasks performed by the underlying detection machinery comprise stimulus detection, signal processing (including amplification and integration of sensory inputs) and production of appropriate responses. In the signal transduction defined as such, two processes predominate: transmembrane signal transfer and sensing of the environment inside the cell. In the first process, a transmembrane sensor converts an extracellular stimulus into an intracellular response. The second process requires that the environmental conditions 'enter' the cell, via molecules diffusing in passively or via permeases. In addition, physical properties, such as light and temperature, are not halted by the cell boundary, and can be perceived in the cell interior.

Proteins make up the sensing and signaling machinery. In bacteria signaling proteins are built from modular components that regulate input, output and protein-protein communication. Many signaling proteins contain characteristic transmitter and receiver domains that promote information transfer within and between proteins. These domains function in conjunction with a variety of (input) sensor domains and (output) effector domains. Signaling pathways are assembled by arranging these domains in various configurations [1].

The simplest circuits have two protein components: a sensor monitoring an environmental parameter, often located in the cytoplasmic membrane, and a cytoplasmic response regulator that mediates an adaptive response (*i.e.* a change in gene expression). Such two-component systems commonly occur in bacteria [1]. A recent analysis of prokaryotic genomes has revealed that the majority of signal transduction systems consist of a single protein with input and output domains, but lacking phosphotransfer domains, as typical for two-component systems. Such one-component systems are evolutionary older and more widely distributed among prokaryotes [2].

Sensors typically contain an N-terminal input domain coupled to a C-terminal transmitter module. Response regulators typically comprise an N-terminal receiver module connected to one or more C-terminal output domains. Detection of a stimulus leads to the modulation of the signaling activity of its associated transmitter to communicate with its corresponding response regulator. The receiver domain of the response regulator detects the incoming signal and alters the activity of its associated output domain to trigger the response [3].

A widespread mechanism for transmitter-receiver communication involves phosphorylation and dephosphorylation reactions. Transmitters have an autokinase activity that reversibly phosphorylates a histidine residue using phosphoryl groups of ATP. Some transmitters also have phosphatase activity toward their receivers. The product phosphohistidine serves as a

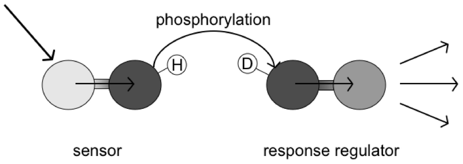


Figure 1.2: **Schematic representation of two-component signal transduction** The sensor contains an input domain (light grey) and a transmitter domain (dark grey). The response regulator contains a receiver domain (dark grey) and an output domain (light grey).

high energy intermediate for transfer of the phosphoryl group to an aspartate side chain in the receiver. The receiver catalyzes the transfer, using the phosphohistidine as substrate. Receivers also catalyze the hydrolytic loss of their phosphate groups [4,5].

Usually, N-terminal transmembrane α-helices position the transmitter containing proteins at the cytoplasmic membrane, with the transmitters projecting into the cell. Consequently, the input domain can reside in the cytoplasm or in the periplasmic space between the inner membrane and the cell wall. The primary structure of input domains differs broadly, reflecting the wide variety of physical and chemical stimuli they detect. Communication with the transmitter domain occurs via stimulus-induced conformational changes of the linker regions. Receiver domains are generally cytoplasmic and may have DNA binding or other regulatory functions. The receiver domain is linked to its output domain(s) via flexible linkers, implying that the receiver requires flexibility to exert control over its adjoining output domain. When assembling signaling pathways, transmitters and receivers are well suited as circuit elements [1,5,6].

Signal transduction in higher organisms

In multi-cellular organisms, single cells also perceive changes in their environmental conditions, in spite of the involvement of extensive homeostatic mechanisms. Several important physiological responses, including vision, smell and stress response involve large metabolic effects produced from a small number of input signals, such as light and the concentration of chemical compounds. The latter includes hormones, growth factors and neurotransmitters, facilitating communication between cells.

There are three major classes of signal transducing receptor proteins, all located at the cell membrane. The classes comprise ion channel-linked proteins 1.3(a), enzyme-linked receptors 1.3(b), and receptors linked to a G protein 1.3(c). Receptors linked to ion channels govern their opening and closing. The latter two classes of receptor proteins comprise three parts. The extracellular domain receives and recognizes a specific signal, whereas the transmembrane domain transmits the signal into the cell. Subsequently, the intracellular domain elicits a response, and in many cases, amplifies the signal through a network of intracellular pathways.

As figure 1.3(b) displays, G proteins coupled receptors are heterotrimeric complexes, comprising an α, a β and a γ subunit, with α binding a guanine nucleotide). Activation occurs via

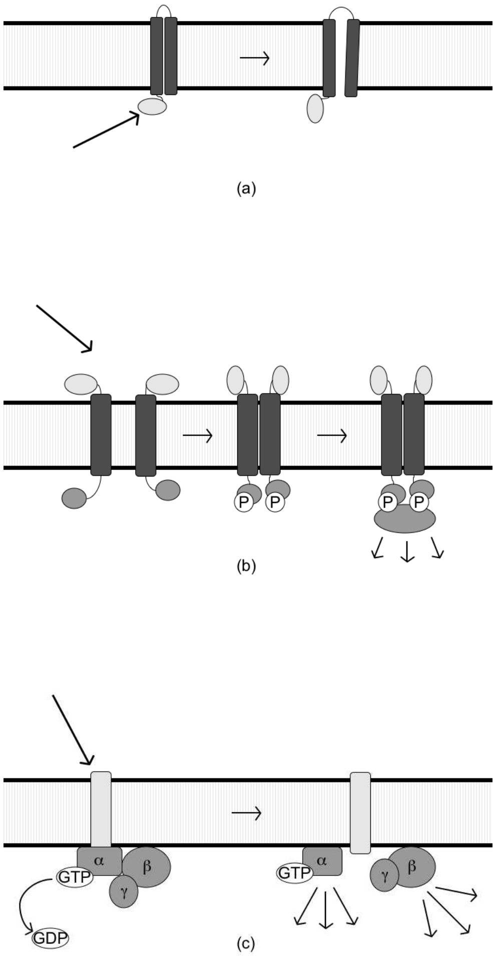


Figure 1.3: **Schematic representation of signal transduction in eukaryotic cells.** (a) Ion channel (b) enzyme linked receptor (c) G protein linked receptor. The thick arrows indicate a signal, perceived by the light grey colored sensor domains. The dark grey colored domains are the signal transduction domains that communicate the signal further into the cell, as indicated by the thin arrows.

binding of guanine-tri-phosphate (GTP), in exchange with GDP. Upon GTP activation, the G protein complex dissociates into GTP-G_α and $G_{\beta\gamma}$, that both are able to activate effector proteins. The hydrolysis of GTP into GDP inactivates the G protein. For example, rhodopsin binds to the G protein transducin, when it is activated by light. Activated rhodopsin catalyzes the exchange of GTP for bound GDP in its cognate G protein, facilitating the activation of a phosphodiesterase with great catalytic prowess via the release of the G_α subunit. This process allows the generation of a nerve impulse by one single photon.

Receptors for growth hormones are an example of enzyme-linked receptors, as they are linked to tyrosine kinases. Upon binding of the hormone, the receptor proteins dimerize. Formation of the dimer induces autophosphorylation, and subsequently the binding of target proteins to the receptor dimers, see figure 1.3(b) for a schematic representation. The target proteins recognize the activated receptors through SH domains: these domains have a high affinity for phosphorylated peptides. Further transduction of the signal into the cell occurs through a cascade of activated tyrosine kinases [7,8].

Sensor characteristics

The biological properties of a protein molecule depend on its interaction with other molecules, in particular specific ligands [9]. This is true for all proteins, including receptors. Sensor proteins have two types of interactions: signal perception and signal communication.

How does a protein perceive a signal? The answer to this question depends strongly on the type of signal. Small molecules, such as a short peptide, a nucleotide or an ion, can bind to a specific region on the sensor protein. The specific amino acids present in this region, the binding site, strongly influence the affinity of the sensor protein for a molecular signal. Incorporation of a co-factor in the sensor protein extends the sensing capabilities of proteins. For example, a heme-cofactor embedded within a protein facilitates the binding of diatomic gases, such as oxygen, carbon monoxide or nitric oxide. Amino acids in the heme binding site tune the affinity of the protein for these molecules, and thus facilitate discrimination between diatomic gases [10]. Photoreceptors provide another example: these proteins contain a light-sensitive cofactor that undergoes a chemical reaction, activated by photon absorption. The chemical structure of the prosthetic group, in combination with the surrounding amino acids in the protein, determine the wavelength of the light that activates the receptor. The perception of another physical property, temperature, can be based on an interplay between the fluidity of the cell membrane and proteins attached to it [11]. In all these sensing mechanisms the structure and chemical properties of the protein determine its specificity for a particular signal.

Following the occurrence of a perception event is the transduction of the signal further into the cell. Binding of a signaling molecule to the sensor, or light-activation of a co-factor, can induce a conformational rearrangement, facilitating a change in the interactions with other proteins (e.g. kinases). As an example, the bacterium *Sinorhizobium meliloti*, a symbiont of rice, contains a cascade to regulate the expression of nitrogen fixation genes in low oxygen concentrations. FixL is the heme-based sensor kinase and FixJ its regulatory partner. In the absence of O_2 FixL catalyzes the phosphorylation of FixJ, thus facilitating the transcription of the nitrogen fixation genes at low oxygen concentration. Binding of oxygen switches off the kinase activity of FixL. FixJ has several roles in this cascade: it serves as the substrate for the second phosphoryl transfer and tunes the affinity of FixL for oxygen [12, 13].

The sensing domain of FixL is a member of a family of protein domains involved in signal transduction and mediation of protein-protein interactions. This family is called PAS after the three proteins first discovered to exhibit its sequence features: Per, involved in the regulation of circadian rhythms; ARNT, functioning as hydrocarbon detector and Sim, a developmental regulator [14]. After this initial identification, many additional PAS domains were identified, in all kingdoms of life. A human PAS domain was discovered in the sequence of a potassium channel called HERG (human ether-a-go-go related gene) [15] where it controls the current of K^+ through this channel via intra-protein interactions [16]. Mutations in HERG can cause a prolongation of cardiac repolarization, the clinical hallmark of the long QT syndrome, an inherited disorder with a propensity for syncope and arrhythmic sudden death. Alterations in the PAS domain can lead to defective trafficking of the protein to the cell membrane [17] and acceleration of channel deactivation [18, 19].

The example of FixL illustrates that PAS domains can function as input domain in receptor proteins, whereas HERG shows that PAS domains also mediate interactions between protein domains. As sensory domains, they detect a wide variety of signals, including light, oxygen and small organic compounds. Currently, many structures of PAS domains are available, from crystallography, NMR spectroscopy and prediction methods. As it is easily accessible for analysis by experiments, the Photoactive Yellow Protein (PYP) from *Halorhodospira halophila* serves as the structural prototype for the PAS family [20]. This protein contains a chromophore, para-hydroxy-coumaric acid, that is covalently linked to its unique cysteine. Absorption of a blue light photon, triggers a series of events occurring at various time scales and encompassing different parts of the protein, resulting of the formation of a partially unfolded signaling state. On a sub-second time scale the protein thermally relaxes back to the resting state. As the signal (*i.e.* light) triggers a global conformational change, perception of the signal in PYP is closely related to the communication of the signal.

Outline of this thesis

Similarities in amino acid sequences has led to the discovery of the PAS family. When crystal structures of PAS domains became available, their structural similarity was striking. This observation has led to the research presented in Chapter 2, presenting the crystal structure of a minimal PAS fold. The design of the minimal PAS fold is based on the structural prototype of the PAS family, Photoactive Yellow Protein (PYP), via truncation of N-terminal residues that are not part of the PAS core. Using structure based sequence alignments, we further defined the PAS core. This definition, combined with structural ensembles generated from distance constraints taken from a crystal structure, served as input in Essential Dynamics analyses. These analyses identify collective motions in protein structural ensembles. The definition of the PAS core enabled quantitative comparison of the PAS domain flexibilities, leading to the conclusion that PAS domains contain similar flexible regions.

The main goal of the work presented in this thesis is to provide a better understanding of the mechanisms underlying signal perception and transduction in PAS domains. In PYP these processes have been the subject of many investigations, leading to a detailed description of the events taking place during its signal transduction cycle. Due to the large range of time scales involved, some issues required advanced molecular dynamics techniques to understand the molecular mechanism underlying the light-triggered changes. Chapters 3 and 4 present parallel tempering simulations of PYP in several states of its signal transduction cycle. Chapter 3 describes the prediction of the structure of the signaling state of PYP, while Chapter 4 shows the conformational requirements for a fast recovery of the resting state.

The comparison performed in Chapter 2 requires the different protein structural ensembles to have identical size. To achieve this, protein parts not belonging to the PAS definition were omitted in the analysis. Chapter 5 presents a new method for the analysis of intramolecular protein motion, called Protein Triads. After an explanation of the method and comparison with the well-known Essential Dynamics technique, both methods are applied to a simple system of eleven atoms moving in a hinge-bending motion, and a parallel tempering simulation of Photoactive Yellow Protein. This chapter ends with Protein Triads analyses on two more PAS domains, FixL and HERG, and shows that their dynamics differ.

Chapters 2-5 all investigate PAS domains in isolation. Chapter 6 describes molecular dynamics simulations of TraR, a quorum sensor from *Agrobacterium tumefaciens*. The Tra quorum sensing system represents a one-component signaling pathway [2] with a PAS domain as input domain. Bacteria use chemical signals to regulate gene expression at high population densities. These cell-cell communication systems require the release and detection of molecules that can diffuse through the plasma membrane. Receptor proteins sense the concentration of these molecules: when this concentration is above a certain threshold level activation of a specific set of genes occurs. After sensing, the receptors activate or repress transcription of genes. These genes may code for proteins involved in various functions, including pathogenesis, sporulation and gene transfer. This process is referred to as quorum sensing. Our simulations show that the auto-inducer enhances fluctuations in the protein that facilitate DNA-binding.

Since the discovery of the PAS family, knowledge on the PAS fold has expanded widely. Chapter 7 gives an overview of the current literature on PAS domains. In addition, two appendices provide general background information on protein structure (appendix A) and molecular dynamics (appendix B), respectively.

Chapter 2

Investigating the structure and dynamics of PAS domains*

PAS (PER-ARNT-SIM) domains are a family of sensor protein domains involved in signal transduction in a wide range of organisms. Recent structural studies have revealed that these domains contain a structurally conserved alpha/beta-fold, whereas almost no conservation is observed at the amino acid sequence level. The photoactive yellow protein, a bacterial light sensor, has been proposed as the PAS structural prototype yet contains an N-terminal helix-turn-helix motif not found in other PAS domains. Here we describe the atomic resolution structure of a photoactive yellow protein deletion mutant lacking this motif, revealing that the PAS domain is indeed able to fold independently and is not affected by the removal of these residues. Computer simulations of currently known PAS domain structures reveal that these domains are not only structurally conserved but are also similar in their conformational flexibilities. The observed motions point to a possible common mechanism for communicating ligand binding/activation to downstream transducer proteins.

Introduction

PAS domains are structural modules that are found in proteins in all kingdoms of life [21, 22]. The PAS module was first identified in the *Drosophila* clock protein PER and the basic helix-turn-helix containing transcription factors ARNT (aryl hydrocarbon receptor translocator) in mammals and SIM (single minded protein) in insects [23]. Most PAS domains are sensory modules, typically sensing oxygen tension, redox potential or light intensity [20, 21]. Alternatively, they mediate protein-protein interactions or bind small ligands [24]. Although the amino acid sequences of the different PAS domains show little similarity, their three-dimensional structures appear to be conserved. All PAS domains resemble the structure of photoactive yellow protein (PYP) [20], a photoreceptor involved in a phototactic response of the bacterium *Halorhodospira halophila* to blue light [25]. Its structure reveals an α-β-fold with the light sensitive chromophore p-coumaric acid bound to the protein via a thioester linkage [26]. Many investigations, including molecular simulation, have elucidated the catalytic function for this protein, *i.e.* signal genera-

---

*This chapter is based on J. Vreede, M.A. van der Horst, K.J. Hellingwerf, W. Crielaard and D.M.F. van Aalten *PAS domains - Common structure and common flexibility?* J. Biol. Chem. 2003, 278:18434-18439 tion and transduction. Upon absorption of a blue light photon, the protein undergoes a photo cycle, linked to the isomerisation and protonation of the chromophore [27–31].

During the photo cycle distinct conformational changes occur that may translate the photonic signal into a cellular response via subsequent protein-protein interactions. Results from early molecular dynamics studies suggested that concerted motions, linked to the chemical state of the chromophore are present in the ground state and that these motions are amplified upon activation (by isomerisation and protonation) of the chromophore [32, 33]. In characterizing these motions, conserved glycines appeared to act as hinges, allowing sub domains in the protein to fluctuate relative to each other. Altering these residues, and so the rigidity of the backbone confirmed their importance for PYP signal transduction. These glycines fall within the PAS fold, and moreover, show a large degree of conservation throughout the PAS family [34]. Such a strong conservation in a protein family with so little similarity at sequence level implies that not only the structure of the PAS family is conserved, but also that these domains have similar conformational freedom.

Here we investigate whether the dynamics properties of PAS domains are intrinsic and associated with their conserved fold. First, we have mutated PYP into a minimal PAS fold, by removing the N-terminal cap (residue 1-25) [35]. Crystals of this mutant resulted in a structure at 1.14 Å resolution that is used in a comparative computational investigation of the conformational flexibility of four PAS domain structures: HERG, the N-terminal domain of a human potassium channel [15], FixL, a bacterial oxygen sensor [36], LOV2 a photoreceptor domain in fern [37] and wt-PYP. Essential dynamics analyses on the sampled configurational space revealed conserved concerted motions. The common structure of PAS domains implies common flexibility, a conserved property fundamental for PAS domain signal transduction.

Materials and methods

Crystallization, diffraction and refinement

Δ_{25}PYP encompassing residues 26-125 of PYP were over-expressed and purified as described elsewhere [35]. Equilibration of 1 μl of 30 mg/ml protein with 1 μl of mother liquor (1.8 M ammonium sulfate, 1o mM $CoCl_2$, 100 mM MES, pH 6.5) against a 1 ml reservoir of mother liquor resulted in crystals after 2-3 days, with a largest dimension of 0.4 mm. Diffraction data were collected at beam line ID14-EH1 (European Synchrotron Radiation Facility, Grenoble, France) and processed with the HKL package (table 2.1) [38]. The structure of Δ_{25}PYP was solved by molecular replacement with AMoRe [39] using the native PYP structure (2PHY) [26] as a search model, excluding the chromophore against 4-8 Å data. A solution was found (r = 0.479, correlation coefficient = 0.282 with two molecules in the asymmetric unit. Initial refinement was carried out with CNS [40] interspersed with with model building in O [41]. The chromophore was not included in the refinement until it was well defined by an unbiased $F_o - F_c$, ϕ_{calc} map (see Fig. 2.1). Further rounds of refinement with SHELX97 [42] allowed the placement of water molecules and the assignment of some alternate conformations. In the final stages of refinement hydrogen atoms were included.

Residues 113 and 114 in one monomer and residue 116 in the other monomer were disordered, although some evidence for several possible conformations was visible in the map. Al-

Cell dimensions	a = b = 82.566 Å, c = 63.453 Å
Resolution range Å	17-1.14 (1.18-1.14)
No. observed reflections	299780 (23225)
No. unique reflections	75208 (7010)
Redundancy	4.0 (3.3)
$\frac{I}{\sigma I}$	13.5 (4.3)
Completeness (%)	94.4 (89.2)
R_{merge}	0.060 (0.217)
R, R_{free}	0.147, 0.177
No. groups	200 residues, 406 H_2O
root mean square deviation from ideal geometry	
Bonds(Å)	0.010
Angles(°)	2.0
B-factor root mean square deviation (Å^2)	
all bonds	2.6
B (Å^2)	15.8 (protein), 33.5 (water)

Table 2.1: **Details of data collection and refinement.** Values between brackets are for the highest resolution shell. Crystals were of space group $P4_32_12$ (a = b = 82.57 Å, c = 63.45 Å) and were cryo-cooled to 100 K. All measured data were included in structure refinement.

though building these regions was attempted, their conformation could not be determined with confidence. PYP in the $P6_5$ space group [33,34,43] suffered similar problems. At the N-terminus, Ala-27 could be modeled in well-defined electron density at the early stages of refinement. Subsequent maps also defined the conformation of Leu-26.

CONCOORD simulations

To sample the conformational space of systems of interest we used the CONCOORD method [44]. In short, CONCOORD extracts distance constraints from a structure, followed by fitting randomly generated coordinates within these contraints. The systems of interest are four PAS domains; HERG, PYP, FixL and LOV2, as well as Δ_{25}PYP and lysozyme. Lysozyme is included as a negative control, since its structure bears no resemblance to the PAS fold. For each system distance constraints were generated from their respective crystal structures (PDB codes: HERG - 1BYW, PYP - 2PHY, FixL - 1DRM, LOV2 - 1G28, lysozyme - 135L, and for Δ_{25}PYP - the structure described here). Subsequently, 1000 structures were generated fitting these constraints, using a damping factor of 0.25 to avoid unreasonable side chain geometries.

Essential dynamics

Essential dynamics determines concerted atomic motions from an ensemble of structures [45], here the CONCOORD trajectories. To describe the correlation of positional shifts of one atom relative to other atoms, the method starts with the construction of a covariance matrix according to Eq. 2.1:

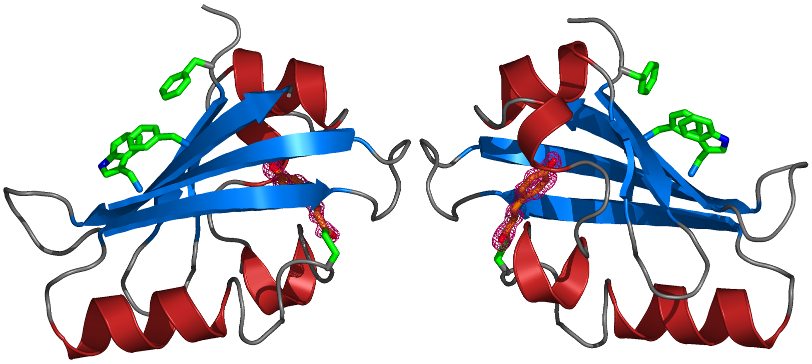


Figure 2.1: **Ribbon representation of the crystal structure of Δ_{25}PYP.** The asymmetric unit cell contains two proteins. Secondary structure elements are marked in the structures. The $F_o - F_c$,Φ_{calc} map just before including the chromophore is shown in magenta, contoured at 2.5 σ. Hydrophobic residues that have become solvent-exposed because of the deletion of residues 1-25 are shown as green sticks.

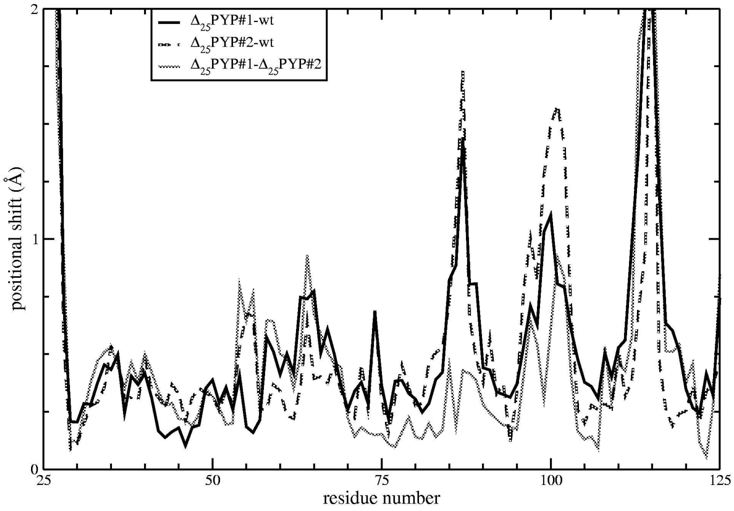


Figure 2.2: **Conformational changes.** Positional shifts of equivalent C_α atoms after superposition of the two Δ_{25}PYP monomers in the asymmetric unit on wtPYP and each other.

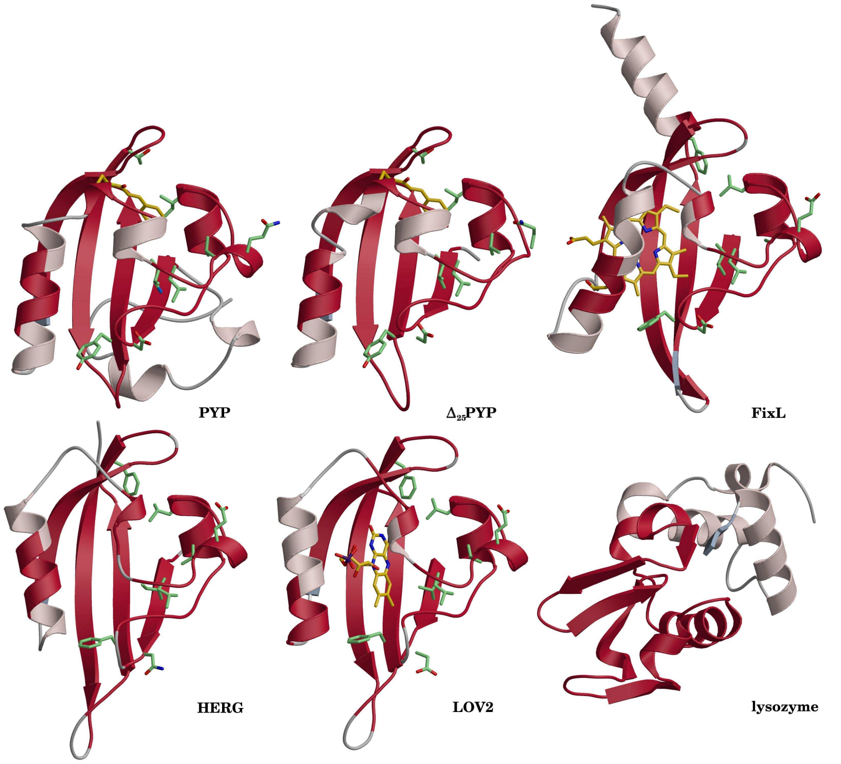


Figure 2.3: **Comparison of the PAS cores.** Ribbon representations of the structures of wtPYP, the Δ_{25}PYP mutant described here, FixL, HERG, LOV2 and turkey lysozyme. Residues that align structurally and are used for comparisons are highlighted in red. The cofactors, if present, are drawn in yellow colored stick model. Homologous residues at similar structural locations (Fig. 2.4) are depicted in green.

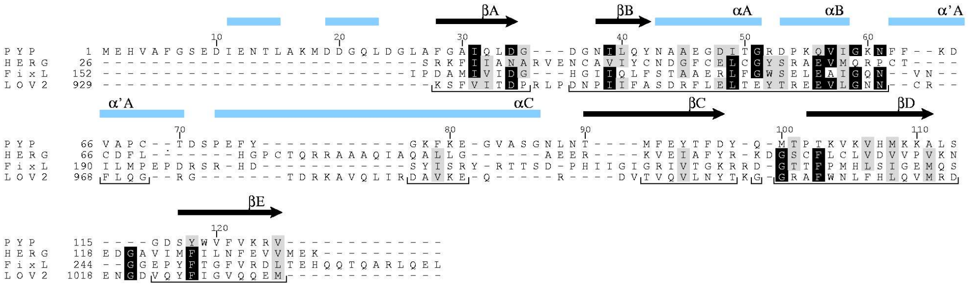

βA
βB
αA
αB
α'A
10
20
30
40
50
60
PYP 1 MEHVAFGSEDIENTLAKMDDGQLDGLAFGAIQLDG---DGNILQYNAAEGDITGRDPKQVIGKNFF--KD
HERG 26 --------------------------SRKFIIANARVENCAVIYCNDGFCELCGYSRAEVMQRPCT----
FixL 152 -------------------------IPDAMIVIDG---HGIIQLFSTAAERLFGWSELEAIGQN--VN--
LOV2 929 ---------------------------KSFVITDPRLPDNPIIFASDRFLELTEYTREEVLGNN--CR--
α'A
αC
βC
βD
70
80
90
100
110
PYP 66 VAPC--TDSPEFY------------GKFKE-GVASGNLNT----MFEYTFDY-Q-MTPTKVKVHMKKALS
HERG 66 CDFL-------HGPCTQRRAAAQIAQALLG------AEER----KVEIAFYR-KDGSCFLCLVDVVPVKN
FixL 190 ILMPEPDRSR-HD------------SYISRY-RTTSD-PHIIGIGRIVTGKRRD-GTTFPMHLSIGEMQS
LOV2 968 FLQG--RG-------TDRKAVQLIRDAVKE-Q------R----DVTVQVLNYTKG-GRAFWNLFHLQVMRD
βE
120
PYP 115 ----GDSYWVFVKRV-------------
HERG 118 EDGAVIMFILNFEVVMEK----------
FixL 244 --GGEPYFTGFVRDLTEHQQTQARLQEL
LOV2 1018 ENGDVQYFIGVQQEM-------------

Figure 2.4: **Structure-based sequence alignment.** Alignment based on the superposition of PYP, HERG, FixL and LOV2 using DALI and WHAT IF. Black arrows indicate β-strands, grey bars indicate α-helices, and labels identify the secondary structure elements. Residues selected for essential dynamics analysis are underlined. Homologous residues in the alignment are colored black for at least three identical residues and colored in grey for at least three homologous residues.

$$C_{ij} = ((x_i - x_{i,0})(x_j - x_{j,0})) \tag{2.1}$$

where x_i and x_j represent the coordinates of atoms i and j in a conformation of the ensemble, and $x_{i,0}$ and $x_{j,0}$ represent the atomic coordinates averaged over the ensemble. The average is calculated over all structures after translational and rotational superimposure on a reference structure. Diagonalizing the covariance matrix yields a set of eigenvectors and eigenvalues. The eigenvectors are directions in a $3N$ dimensional space, with N the number of atoms in the analysis. These directions represent concerted displacements of groups of atoms in Cartesian space. The eigenvalues are a measure of the mean square fluctuation of the system along the corresponding eigenvectors and their size determines the order of eigenvalues: the first eigenvector has the largest eigenvalue. Previous investigation has shown that only a small percentage of all eigenvectors involve large concerted motions, leading to the selection of essential eigenvectors and so the definition of an essential subspace [46].

To compare the essential subspaces of the systems under investigation, the dimensions must be identical in size, *i.e.* each system must contain an identical number of atoms. First, only C_α atoms are included [46] and second, the PAS fold is reduced to a common core. The DALI server pairwise compares structures and aligns them on their secondary structure content and sequence [47]. Combining these pairwise alignments yielded the common elements present in all PAS structures. The common core residues were used for essential dynamics analysis for each PAS structure. Starting from the N-terminus, an equal number of residues was selected for lysozyme.

Results and discussion

The crystal structure of Δ_{25}PYP at atomic resolution

The structure of Δ_{25}PYP was solved by molecular replacement and refined to a 1.14 Å resolution (R-factor = 0.147, R_{free} = 0.177). The asymmetric unit contains two protein molecules related by a non-crystallographic 2-fold rotation axis. The molecules have a similar conformation with a root mean square deviation of 0.77 Å on the position of C_α atoms. Compared with wild type PYP, the two molecules superimpose with rms deviations of 0.99 and 0.76 Å respectively.

Fig. 2.2 shows the positional shifts of the C_α in these superpositions. Residues 26-27, 84-88, 98-101 and 111-117 show the largest deviations with the wild type. Deletion of the N-terminal 25 residues explains the deviation at residues 26-27. Excluding these residues from the superposition, the rms deviation from wild type is decreased with 0.2 Å. Previous structural investigations, NMR and PYP in different crystal space groups, have shown that the loop around Met-100 is flexible. Close contacts between the monomers in the asymmetric unit cell affect the conformation of this "100-loop", and indeed, the distance between the two methionines is smaller than 4.0 Å. A similar explanation based on crystal contacts holds for the deviations in the loops comprising residues 80-84 and 111-117. The latter loop contains disordered residues, that might contribute to the observed flexibility in this loop. Native PYP contains two hydrophobic cores, one within the PAS domain, located between the β-sheet and the αC helix, and another between the β-sheet and the two small N-terminal helices. In the latter core, residues Phe-28 (βA), Trp-

99 and Phe-119 (both on βE) extend toward the N-terminal domain and are solvent exposed in Δ_{25}PYP.

No dimerization through crystal contacts seems to occur, agreeing with the observation that the fluorescence emission of Trp-119 is enhanced and blue-shifted [48]. The environment of these residues is more polar, likely causing the decrease in thermal stability of Δ_{25}PYP. In summary, the removal of the first 25 residues of PYP does not significantly affect the overall fold of the PAS core. In agreement of these observations, spectrophotometric data on the mutant show a minimal blue-shift for the absorbance maximum of the chromophore, indicating little perturbation in the direct chromophore environment. Also, similar photocycle intermediates as in the wild type are present. Exposure of several hydrophobic residues in the structure of Δ_{25}PYP may explain the slower recovery observed in the photo cycle kinetics of the mutant.

PAS domain flexibility

The common fold of PAS domains, in combination with the similarity in function leads to the hypothesis that these domains share common dynamic properties. These shared flexible modes might allow the domains to communicate with signal transduction partners through a conserved mechanism. Here, we have sampled the conformational space of five PAS domains, including the minimal PAS fold Δ_{25}PYP. Using the CONCOORD sampling method, structural ensembles were generated initiated from the structures at atomic detail. However, only C_α atoms that are part of the common PAS core are included in the subsequent essential dynamics analysis (Fig. 2.3).

Structure-based sequential alignment defined 78 residues as part of the common PAS core, including most of the central β-sheet, helices αA/B and part of the long αC helix. The essential dynamics analyses yielded sets of eigenvectors (flexible modes) for each structural ensemble, sorted by eigenvalue size. The first 5 % of the eigenvectors (12) describe more than 95 % of the motions in the sampled conformational spaces. This condensed description of the flexibility facilitates comparison of the dynamic properties of PAS domains.

The selected eigenvectors span an essential subspace. A measure of similarity of these subspaces is the cumulative square inner product of the eigenvector sets, since the number of atoms included in the essential dynamics analyses is equal for all the systems of interest. From table 2.2, that lists the values resulting from pairwise comparison of the subspaces, it is clear that the sets of essential eigenvectors of the PAS domains investigated here are very similar. This observation suggests that the PAS cores share common motions that are not present in lysozyme (included here as a negative control). Projection of the PAS eigenvectors onto the first three eigenvectors from the wt-PYP ensemble allows further confirmation, Fig. 2.5. The sets of essential eigenvectors reproduce the wt-PYP eigenvectors for 90 %, whereas the eigenvectors resulting from sampling the conformational space of lysozyme reproduce only 50 % of a PYP flexible mode at the highest. Thus, PAS domains share common flexible modes in addition to a common structure.

	PYP	FixL	HERG	LOV32	Δ_{25}PYP	Lysozyme
PYP	1.00					
FixL	0.70	1.00				
HERG	0.66	0.72	1.00			
LOV2	0.69	0.71	0.73	1.00		
Δ_{25}PYP	0.78	0.68	0.69	0.70	1.00	
Lysozyme	0.24	0.24	0.22	0.24	0.23	1.00

Table 2.2: **Comparison of essential spaces.** The first twelve (*i.e.* 5 % of the total dimension of the system) eigenvectors are pairwise compared through calculation of a cumulative square inner product.

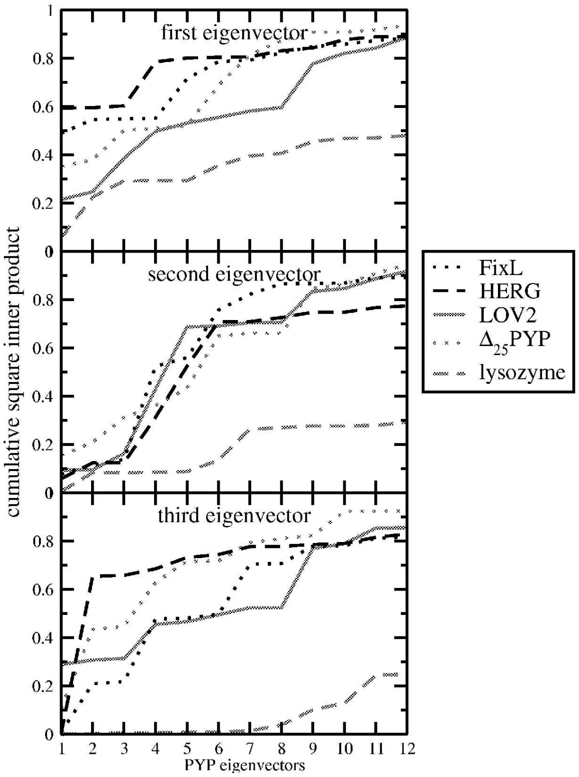


Figure 2.5: **Comparison of PAS CONCOORD eigenvectors.** Cumulative mean squared inner products of the first, second and third eigenvectors obtained from the CONCOORD ensembles of FixL, HERG, LOV2, Δ_{25}PYP and lysozyme against the first 12 eigenvectors obtained from the wt-PYP CONCOORD ensemble.

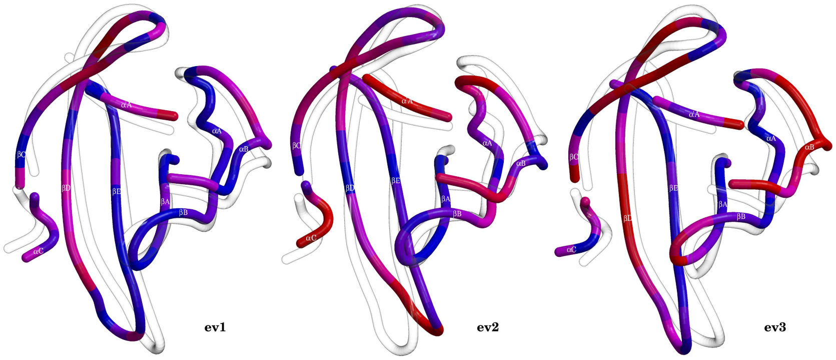


Figure 2.6: **PYP conformational changes.** Atomic positional shifts as described by the first three eigenvectors of Δ_{25}PYP are depicted as structures in Cartesian space. The projection -2 nm along the eigenvectors is colored, and at +2 nm the projection is transparent. Relative degrees of positional shifts are indicated from blue (smallest fluctuations) to red (largest fluctuations). the transparent structures indicate the direction of the motion.

Now that we have established that PAS domains share common flexible properties, the nature of these motions is the next issue to address. To understand the essential subspaces at a molecular level, the underlying eigenvectors are translated to atomic positional shifts in Cartesian space, at -2 nm and +2 nm along a particular eigenvector. Fig 2.6 shows the conformations resulting from such a translation of the first three eigenvectors from wt-PYP. Changes in positional shifts colour the conformation projected at -2 nm. The central β-sheet is static in comparison to the loops, most notably the αA/αB segment. The latter region is generally important for ligand bindipyp2ng. In PYP, the co-factor is bound via a cysteine and a glutamate, whereas in FixL and LOV2 a phenylalanine in similar position interacts with the heme group and FMN respectively. Also, this region contains a saltbridge, conserved in PAS domains, and implicated in signal transduction [49]. In this segment of PAS structure, similarity at amino acid level is very low.

Concluding remarks

PYP maintains the PAS fold, even in absence of residues that do not belong to the PAS core, despite the exposure of several hydrophobic residues to solvent. The structure of this Δ_{25}PYP, combined with structures of other PAS domains allowed further comparison of the PAS family. Although these domains share little similarity at sequence level, their resemblance in three-dimensional structure is remarkable. Even though these domains bind a variety of ligands, their similar conformations might involve similar conformational changes. We investigated this by sampling the conformational space of PAS structures at atomic detail, followed by essential dynamics analysis of the C_α atoms of the common PAS core. The results show that the αA/B segment moves in a concerted fashion.

Chapter 3

Predicting the signaling state of Photoactive Yellow Protein*

As a bacterial blue light sensor the photoactive yellow protein (PYP) undergoes conformational changes upon signal transduction. The absorption of a photon triggers a series of events that are initially localized around the protein chromophore, extend to encompass the whole protein within microseconds, and lead to the formation of the transient pB signaling state. We study the formation of this signaling state pB by molecular simulation and predict its solution structure. Conventional straightforward molecular dynamics is not able to address this formation process due to the long (microsecond) timescales involved, which are (partially) caused by the presence of free energy barriers between the metastable states. To overcome these barriers, we employed the parallel tempering (or replica exchange) method, thus enabling us to predict qualitatively the formation of the PYP signaling state pB. In contrast to the receptor state pG of PYP, the characteristics of this predicted pB structure include a wide open chromophore binding pocket, with the chromophore and Glu46 fully solvent exposed. In addition, loss of α-helical structure occurs, caused by the opening motion of the chromophore binding pocket and the disruptive interaction of the negatively charged Glu46 with the backbone atoms in the hydrophobic core of the N-terminal cap. Recent NMR experiments agree very well with these predictions.

Introduction

The bacterial blue-light sensor Photo-active Yellow Protein (PYP) originates from *Halorhodospira halophila* [25, 50]. The light sensitive part of PYP is the deprotonated para-hydroxy-coumaric acid, which is covalently attached to the protein via a thio-ester linkage to the unique cysteine at position 69 [51]. The buried negative charge of the chromophore is stabilized in its binding pocket by hydrogen bonds to the surrounding residues Glu46, Tyr42, Thr50 and the protein backbone [26]. Arg52 serves as the lid of the binding pocket, shielding the chromophore from contact with water molecules. Three-dimensional structure analysis identified two basic domains in PYP: a PAS core [20], comprising amino acids 30-125, and an N-terminal cap containing the first 29 residues. The chromophore is contained within the PAS-core [26].

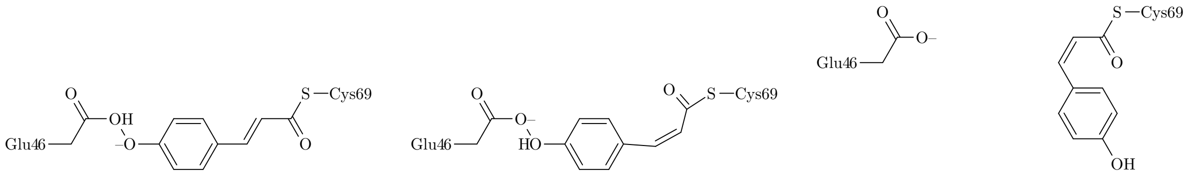
---

[*] This chapter is based on J. Vreede, W. Crielaard, K.J. Hellingwerf and P.G. Bolhuis *Predicting the signaling state of Photoactive Yellow Protein* Biophys. J. 2005, 88:3525-3535

Absorption of a blue photon by the chromophore triggers a sequence of events occurring at various time scales and encompassing different parts of the protein [52]. After excitation the chromophore isomerizes from *trans* to *cis* [29], followed by the disruption of the hydrogen bond between the carbonyl oxygen of the chromophore and the protein backbone [28], resulting in the so-called pR state. Subsequently, on a microsecond time scale, a proton migrates from the Glu46 side chain to the chromophore causing a blue shift in its absorbance maximum [53, 54]. The intermediate associated with this reversible process is denoted as pB′. Further proof of its existence was obtained with the deuterium isotope effect [55]. More recently, resonance Raman spectroscopy has also confirmed the existence of this intermediate [56].

The proton transfer renders the protein metastable by leaving a negative charge at Glu46 and disrupting the stabilizing hydrogen bonding network. The protonation of the chromophore is therefore also the trigger for the formation of pB, a process that occurs on a millisecond time scale [52]. The formation of pB is linked to large conformational rearrangements throughout the protein [57, 58], sometimes even referred to as partial unfolding of the protein [59, 60]. The observation that pB is the longest living state in the photo-cycle, has led to the hypothesis that pB is in fact the signaling state of PYP [52]. The return to the ground state, completing the photo-cycle, is a sub-second process and includes the deprotonation and *cis* to *trans* re-isomerization of the chromophore.

Fig. 3.1 visualizes the chemical structure of the chromophore binding pocket in the three states described above. In pG the chromophore is in a *trans* configuration, deprotonated, and hydrogen bonded to the protonated Glu46. This hydrogen bond is retained in the pB′ configuration, whereas the proton on Glu46 has transferred to the chromophore in *cis*-configuration [61]. The pB state has the same chemical structure, except for the disrupted hydrogen bond between the chromophore and Glu46. The hydrogen bond between the chromophore carbonyl oxygen and the protein backbone is disrupted in pB′ and reformed in pB [56].



(a) pG (b) pB′ (c) pB

Figure 3.1: **Structural differences between the pG, pB′ and the pB state.** In pG the chromophore is deprotonated and in *trans* configuration, hydrogen bonded to the protonated Glu46. In pB′, the chromophore is *cis*-isomerized and the proton has moved to the chromophore, leaving a negative charge on Glu46. The hydrogen bond between the groups is retained, but broken when the protein enters the pB state.

As to the nature and underlying mechanism of the changes upon the formation of pB, two areas in the protein are implicated: the chromophore binding pocket and the N-terminal domain [35]. CD-spectroscopy showed a loss of helical content in the protein upon activation [62, 63], while the change in diffusion constants during the photo-cycle, as measured with transient grating spectroscopy [64], is best explained by conformational rearrangements and the increased exposure of protein hydrophobic interior. Results arising from the use of a hydrophobicity probe agree with the latter, specifying the chromophore binding pocket as the region where the main rearrangements occur [65]. Small angle X-ray scattering experiments showed an increase in the radius of gyration of N-terminal deletion mutants of PYP [66, 67], confining the conformational changes to the PAS core and the first helix of the N-terminal domain. Finally, time resolved fluorescence measurements on the unique tryptophan residue indicate that this residue has varying degrees of solvent contact during the photo-cycle [68].

Time-resolved crystallography on constantly illuminated crystals provided the first three-dimensional model of pB [30]. In contrast to the results described above, this structure only shows differences in side chain orientation in the chromophore binding pocket. More recent results imply, however, that the conformational changes occur throughout the protein [69, 70]. NMR spectroscopy indicated that the formation of a buried negative charge on Glu46 drives the conformational changes in the protein [71, 72].

Molecular simulation methods, such as molecular dynamics (MD) can, in principle, provide a complementary atomistic picture of the changes occurring in PYP during its photo-cycle. For instance, by employing a combination of quantum mechanical and molecular mechanical calculations Groenhof *et al.* recently proposed a model describing the initial events, including the excitation and subsequent rearrangements [73]. Groenhof *et al.* also used parameters from semi-empirical calculations for a protonated chromophore [74], embedded in an equilibrated protein structure, to show initial rearrangements in the chromophore binding pocket and the N-terminal domain [75]. The proton transfer from Glu46 to the chromophore is assigned as the trigger for the conformational change to pB [75]. Simulation data on wild type PYP and a mutant E46Q further elucidate this trigger as the weakening of the hydrogen bond between the chromophore and Glu46 [76].

In the first attempt to model the signaling state pB using MD, the ground state chromophore vinyl bond was replaced with a single bond potential, to allow for faster rearrangements [32]. In another simulation study starting from the crystal structure obtained from illuminated crystals, water molecules enter the chromophore binding pocket and hydrate Glu46 [77]. Both simulations show a slow drift away from the original crystal structure. Unfortunately, conventional MD is limited to relatively small time scales in the order of nanoseconds, while the formation of pB from pB′ is a millisecond process.

A recent study therefore used a coarse-grained Hamiltonian, which resulted in a description of pB as partially unfolded states stabilized by conformational entropy in the N-terminal domain and vibrational entropy around the chromophore [78]. A disadvantage of using such coarse-grained models is that these are not able to resolve the atomic structure of the pB state. In conclusion, although previous studies investigated the initial conditions for pB formation, the extent of the conformational rearrangements and the actual solution structure of pB is still unknown.

The long time scales involved in the formation of pB are partly caused by free energy barriers between the several metastable states. One way to overcome the trapping of biomolecular systems in local minima between these barriers is to perform Parallel Tempering (PT) simulations [79–81]. The PT method, also known as replica exchange, combines multiple molecular dynamics simulations with a temperature exchange Monte Carlo process [82]. The method has been proved useful in folding/unfolding studies on peptides, including α-helices [83], a

β-hairpin [84, 85], protein A [86] and Trp-cage [84, 87]. In this work we employ the parallel tempering technique to overcome the free energy barriers for the formation of pB and study the conformational differences between the receptor- and signaling state of PYP.

Methods

During signal transduction, the Photoactive Yellow Protein (PYP) undergoes conformational transitions toward the signaling state at a microsecond and millisecond time scale. In this work we investigate the formation of this signaling state, by performing five independent parallel tempering (PT) simulations, each based on a different starting structure. The crystal structures of PYP in the dark and the bleached state served as two of the starting configurations, PDB codes 2PHY [26] and 2PYP [30], respectively. Coordinates based on NMR constraints [88] served as input for two more PT simulations. Conformation 11 in the ensemble of solution structures (PDB code 3PHY) has a hydrogen bond between the chromophore and Glu46 and hence was selected as a starting point. As a starting point for the formation of pB, we replaced the original chromophore coordinates in the NMR structure by coordinates from a crystal structure of a cryo-trapped photo-cycle intermediate [89] in *cis* configuration. This replacement did not result in unfavorable atomic interactions. However, the altered protein structure with the *cis* chromophore is not the actual pB signaling state. Instead, this configuration represents the pB′ state in the photo-cycle of PYP. The last of the five PT simulations was initiated from a selected on-way conformation in the nmr pB′ run. This last run was included to speed up the slow equilibration towards the pB state.

Polar and aromatic hydrogen atoms were added to all four starting structures. The ground state pG differs from the signaling state pB not only in conformation (including the *trans* or *cis* configuration of the chromophore), but also in the protonation of the chromophore and Glu46. In pG, the phenolic oxygen of the chromophore is deprotonated, and Glu46 is protonated. Conversely, the chromophore is protonated and Glu46 is deprotonated in the pB state (Fig. 3.1). Aliphatic groups were included as heavy carbon atoms (united atom model). Subsequently, the protein configurations were placed in a periodic dodecahedral box, and immersed in SPC water [90]. The box size included the protein and a radius of 1.5 nm around it. Water molecules that overlapped with the protein or resided in internal hydrophobic cavities were removed. Six water molecules at the most electronegative positions were replaced by sodium ions to neutralize the charge of -6 on the protein. The systems were energy minimized using 200 steps of the conjugate gradient method [91], and equilibrated to dissipate excess energy and relax the box volume. Positions of the water molecules and the hydrogens were relaxed for 10 ps, followed by 100 ps of equilibration of the whole system in the NpT ensemble. The GROMOS force field was used to describe the interactions between the atoms [92–94]. Van der Waals interactions were treated with a cut-off [91] of 1. 4 nm, and PME handled the long range electrostatics [91]. Using constraints, LINCS for interactions between protein atoms [95] and SETTLE for water interactions [96], allowed a time step of 2 fs. Parameters for the chromophore were taken from [75]. Prepared as such, the systems were used as input for the PT simulations.

The GROMACS software package was used for equilibration and parallel tempering, in combination with a PERL script that performed the temperature swaps. The Berendsen thermostat [97], with a coupling constant of 0.1 ps allowed fast adaptation of the systems to temperatures ranging from 280 K to 640 K. Every 1 ps, attempts to exchange temperatures between systems were made. Although the Nosé-Hoover algorithm is in principle the correct thermostat in constant temperature simulations, we found it adjusted too slowly to equilibrate within 1 ps. However, it is not likely that changing the thermostat will alter the qualitative results in this work.

Coordinates prior to every temperature swapping attempt were written out and used for subsequent analysis. The simulations were performed on a home-built Beowulf cluster, using 32 AMD processors each running two replicas simultaneously. The temperatures for the xtal simulations ranged from 283 K to 630 K. In the NMR simulations the temperatures were set between 300 K and 560 K for both the pG and pB′ simulations. The temperatures in the pB parallel tempering run varied between 282 K and 645 K. The temperature gap was initially estimated by a linear dependence on the inverse temperature, and turned out afterward to give rise to a uniform acceptance ratio of around 30%, around the entire temperature domain. After a 2 ns equilibration period, the five independent parallel tempering runs were continued for on average 8-10 ns, amounting to a total simulation time of more than $64 \times 10 \times 5 \approx 3200\mathrm{ns}$.

Various analysis tools included in the GROMACS molecular dynamics package were used here to calculate fluctuations and several order parameters: the distance between the centers of mass between two groups, the number of hydrogen bonds and the radius of gyration [91]. To analyze the extent of solvation in a protein region we subtracted the number of solvent-protein hydrogen bonds $(N_{protein-solvent})$ from (two times) the intra-protein hydrogen bonds $(N_{protein-protein})$ to obtained the hydrogen bond difference parameter:

$$\Delta = 2N_{protein-protein} - N_{protein-solvent} \tag{3.1}$$

This parameter is positive when the protein is not solvated and becomes negative when more water enters the protein region included in the analysis, replacing the intra-protein hydrogen bonds. A donor-acceptor distance less than 0.35 nm and a donor-hydrogen-acceptor angle of less than 60 degrees defined a hydrogen bond. The number of water molecules surrounding a residue and visual inspection in VMD [98] were also part of the analysis. Free energy landscapes or profiles as a function of (a combination of) the above order parameters can give insight into the (meta)stability of PYP. They can be computed by taking the negative natural logarithm of one- or two-dimensional probability histograms, which result from the sampling of the order-parameters at a fixed temperature during the course of a PT simulation.

Results

Using the parallel tempering (PT) technique, we have investigated the functional conformational transitions during the photo-cycle of Photoactive Yellow Protein (PYP). To probe the differences between simulations starting from crystal structures and conformations based on NMR constraints, both the crystal structure 2PHY [26] and a conformation from the nmr solution structure 3PHY [88] served as starting points for simulations of the receptor state pG. To extend this comparison to the signaling state, the crystal structure of a photo-cycle intermediate, 2PYP [30] and a manually altered NMR conformation from 3PHY [88] initiated simulations of the pB state.

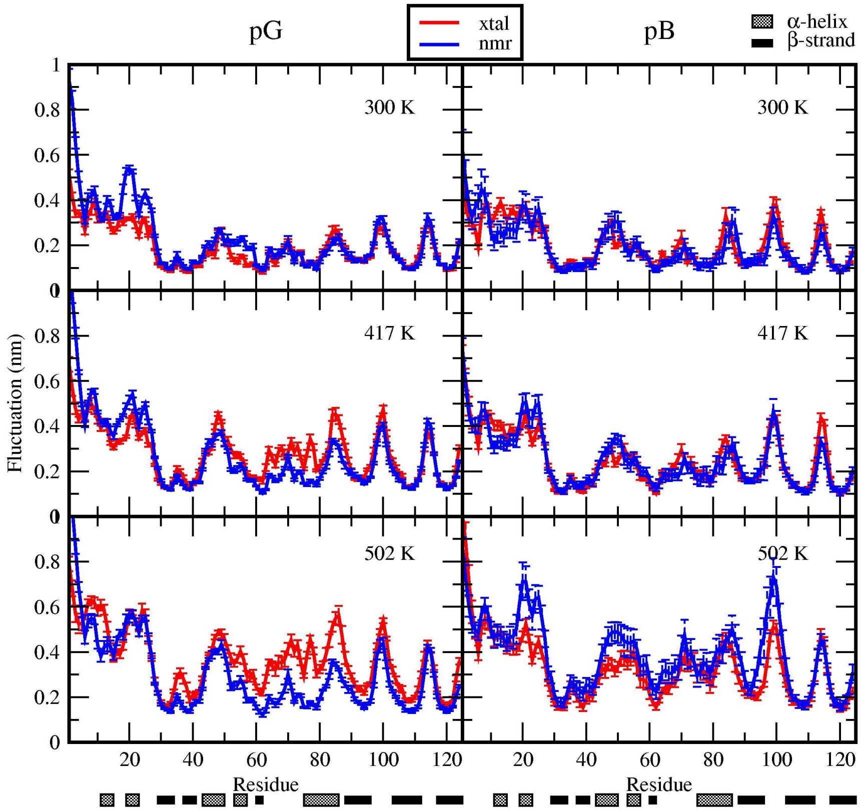


Figure 3.2: **Fluctuations in the protein.** For pG (left panel) and pB (right panel) deviations in atomic displacement are averaged in nm for each residue in PYP at 300 K, 417 K and 510 K. Using fluctuations over 1 ns time intervals, the error bars show the drift during sampling. The labels xtal and nmr indicate the starting structure for the PT. At the bottom, the secondary structure in the protein is indicated by thick, shaded bars for α-helices and smaller, black bars for β-strands.

The label 'xtal' or 'crystal' denotes simulations starting from crystal structure coordinates, and 'nmr' indicates simulations that started with a conformation from a solution structure (see the Methods section for a more detailed description).

Fig. 3.2 shows the root mean square fluctuation (RMSF) in the atomic displacements as a function of the residue number. The values are averaged over all atoms in each residue. The error bars indicate the variance of the fluctuations using simulation blocks of 1 ns. The peaks in the graphs correspond to loops in the protein structure, while the stable parts (below 0.2 nm) correspond to strands of the central β-sheet. With the increase of temperature, the fluctuations in the flexible loops increase also, while the β-strands fluctuate at a value of around 0.2 nm. Additional fluctuation peaks arise around residues in the chromophore binding pocket (CBP), in particular for residues 42-52, 68-72 and 96-100. The first two stretches are part of a helical structure, the last stretch is a loop connecting two β-strands. Higher temperatures cause larger fluctuations in these parts of the protein. The N-terminal domain shows large fluctuations of

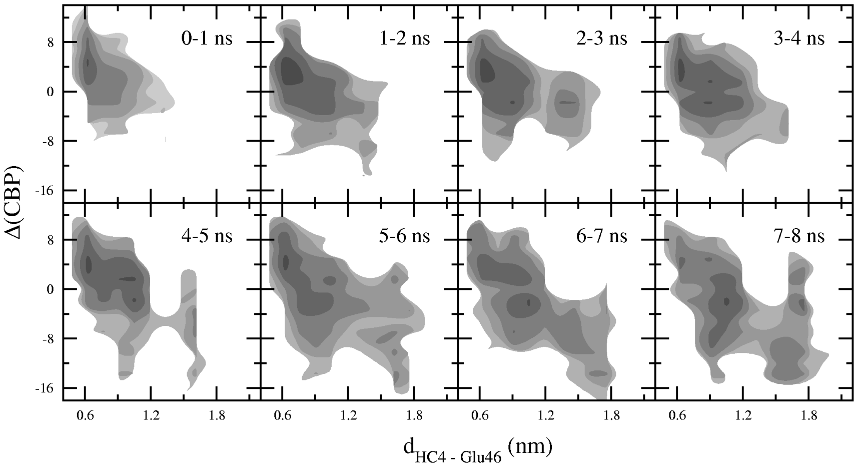


Figure 3.3: **Time evolution of a strained chromophore in the confinement of the protein environment.** The free energy diagrams for the chromophore binding pocket are plotted for each nanosecond of the parallel tempering simulation starting from a protonated *cis*-chromophore [89] in an equilibrated NMR configuration of PYP at 301 K. The free energy is plotted as a function of the distance between the chromophore and Glu46, and the hydrogen bonds difference Δ_{CBP} (Eq. 3.1) in the chromophore binding pocket. The contour lines indicate the k_BT levels, decreasing with darker shading. White areas have not been sampled.

0.5 nm to 1.0 nm for the first two residues, independent of temperature and starting configuration. Regarding the secondary structure in the N-terminal domain, as assigned on basis of the structural elements in the ground state crystal structure [26], the first helix (residues 11-15) is more stable than the second one (residues 19-23). In the nmr simulations of pG and pB the difference in fluctuation between the first and the second helix is around 0.2 nm and at higher temperatures this difference is more pronounced. Unfortunately, the fluctuation graphs do not provide detailed atomistic information on the rearrangements in neither the chromophore binding pocket, nor the N-terminal domain.

The convergence of parallel tempering results starting from entirely different conformations, but with identical chemical structure (*i.e.* simulation topology), would serve as a good test for the quality of the simulation method. Whereas conventional MD simulations are not able to reach global equilibrium for PYP when starting from different conformations, the fluctuations of the crystal PT simulations overlap remarkably well with those of the nmr simulations. This overlap was only achieved after the highest temperature in the crystal PT runs was set to 640 K, whereas the nmr simulations had a maximum temperature of 560 K only. Differences in the interaction between protein and solvent might explain this behavior. Indeed, the potential energy of this non-bonded interaction was around 200 kJ/mol higher for the nmr-based simulations than for the crystal-based simulations.

Focusing on the rearrangements taking place in the chromophore binding pocket, Fig. 3.3 shows the time evolution of the two dimensional free energy diagrams of the nmr-based pB′ simulation as a function of the distance $d_{\mathrm{HC4-Glu46}}$ between the centers of mass of the phe-

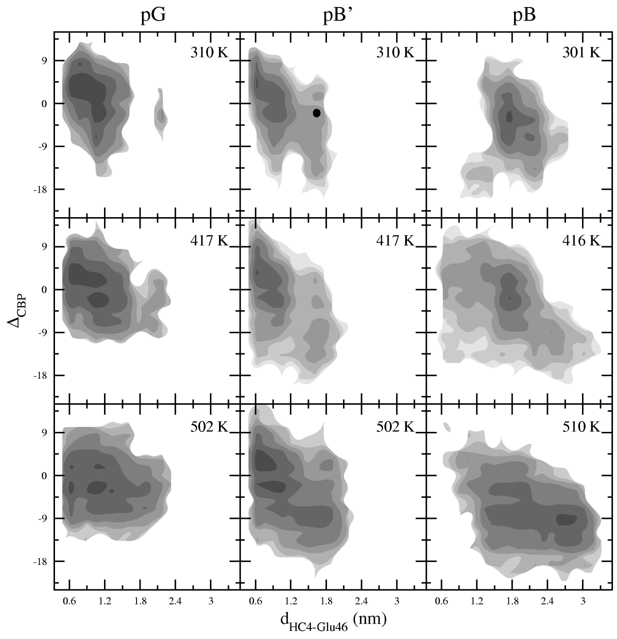


Figure 3.4: **Free energy diagrams** at three different temperatures as a function of the distance between the chromophore and Glu46, and the hydrogen bond difference Δ_{CBP} (Eq. 3.1) in the chromophore binding pocket for PT nmr simulations of the pG, pB′ and pB state, indicated by the labels. The contour lines indicate the k_BT levels, decreasing with darker shading. White areas have not been sampled. A protein conformation within the black area in the 310 frame of the pB′ simulation is selected as a starting point for a new PT run.

nol(ate) ring of the chromophore and the side chain of Glu46, and the hydrogen bond difference Δ_{CBP} in the CBP (for a definition of Δ, see the Methods section) . Included as part of the hydrogen binding pocket are Tyr42, Glu46, Thr50, Cys69, Phe96, Met100 and the chromophore itself. The values of Δ_{CBP} range from +14 in the crystal structure to -24 in the completely solvated pB conformation. The first frame in Fig. 3.3 shows the profile for a chromophore that is buried in the protein and participates in a hydrogen bonding network formed by Tyr42, Glu46 and Thr52. Here, the free energy minimum lies at a Glu46-chromophore distance of 0.63 nm and has a hydrogen bond difference of $\Delta_{\mathrm{CBP}} = 5$. The subsequent time frames show a consistent shift of the sampling toward an increasingly negative value for Δ_{CBP}, combined with a larger distance between Glu46 and the chromophore $d_{\mathrm{HC4-Glu46}}$. A new minimum appears at $d_{\mathrm{HC4-Glu46}}$ = 1.64 nm and Δ_{CBP} = -2, in a region that has fewer intra-CBP-hydrogen bonds and a larger distance between the chromophore and Glu46. Visual inspection shows that the negative charge on Glu46 destabilizes the hydrogen bonded connections and causes the intrusion of water molecules in the protein interior, as indicated by the increasingly negative value for Δ_{CBP}. Ultimately, Glu46 breaks loose from the hydrogen bonding network and becomes exposed to solvent. During the solvent exposure of the chromophore a hydrogen bond forms between the backbone amide of Cys69 and the carbonyl oxygen on the chromophore, stabilizing the solvent oriented conformation. This bond is absent while the chromophore is still buried in the protein.

In Fig. 3.4 the PT free energy of the CBP at three different temperatures is shown for the nmr simulations of the receptor state pG, the initial signaling state pB′ and a final signaling state pB. The complete sampling of the entire configuration space beteen pB′ and pB is too slow, even for the PT simulations. To speed up the sampling a protein configuration from the 310 K run of the pB′ PT simulation that has both the chromophore and Glu46 solvent exposed, was selected as the starting point for a new set of PT replicas. The dot in the 310 K frame of the pB′ simulation indates this chosen configuration and the label pB indicates the new PT simulation run.

In all simulations, the potential energy contributions become less relevant at higher temperatures leading to broader minima and shallower free energy profiles, as the free energy profiles at around 500 K clearly show. The regions sampled in the pG simulation and in the pB′ simulation are similar and show minima around $d_{\mathrm{HC4-Glu46}}$ = 0.63 nm and Δ_{CBP} = 5. In both the pG and the pB′ state the hydrogen bonding network fluctuates between a tightly connected and a more loose structure. The free energy profile of pG at 310 K shows a high barrier that separates a state where the chromophore-Glu46 distance is 1.63 nm at the most, from a state where this distance has a value of 2.19 nm, thus stabilizing the ground state. This free energy barrier disappears at higher temperatures. No barriers larger than a few k_BT are present in the free energy profiles of pB′ at 310 K, and of pB at 301 K, suggesting these states are much less stable. The pB state at 301 K has a different profile in comparison to the pB′ simulation with a second minimum at $d_{\mathrm{HC4-Glu46}}$ = 1.78 nm at Δ_{CBP} = 3. At 416 K the free energy profile of pB extends into two directions, one is a return to the region also sampled by the pB′ simulation, with a small $d_{\mathrm{HC4-Glu46}}$ value and Δ_{CBP} larger than zero, the other samples chromophore-Glu46 distances of above 3 nm and values for Δ_{CBP} that indicate that solvent molecules entered into almost all protein-protein hydrogen bonds.

Fig. 3.5 displays the free energy profiles for the pG and pB simulations initiated from a crystal structure. The profile for the pG state at 302 K shows similar states as those in the profile of the pG nmr simulation at 310 K. The barrier separating the states is less high, and the free energy profile is shallower. At 505 K the sampled region extends to values larger than 3.4 nm. Extension of the PT simulation temperature range to higher temperatures may be the cause for these differences with regard to the nmr simulation of pG. The pB xtal simulation seems to be in a state that lies between the states sampled in the pB′ and pB nmr simulations. At a temperature of 302 K two states occur, at $d_{\mathrm{HC4-Glu46}}$ = 1.15 nm and Δ_{CBP} = 2 and at $d_{\mathrm{HC4-Glu46}}$ = 2.20 nm and Δ_{CBP} = -15. The former is more similar to the pB′ nmr simulation and contains a buried chromophore and Glu46, whereas the latter is closer to the nmr-based pB results, with the chromophore and Glu46 exposed to solvent. At higher temperatures both the free energy profiles resemble those sampled for pB in the nmr simulation.

Upon triggering of the photo-cycle, not only the chromophore containing part in PYP undergoes a conformational transition, but also the N-terminal cap, comprising the first 29 amino acids of PYP, partially unfolds. Its role in the conformational transitions and the degree of unfolding is still unclear, largely because the crystal structure shows more prominent α-helices than the NMR experimental solution structure predicts. One measure for unfolding is the radius of gyration R_{gyr} of the hydrophobic core. Three phenylalanine residues at positions 6, 28 and 121 make up the hydrophobic core in the N-terminal domain. A second order parameter for unfolding is the hydrogen bond difference $\Delta_{\mathrm{N-term}}$ in the helical residues in the N-terminal cap, measuring the solvent exposure. Since some helical residues are always solvent exposed,

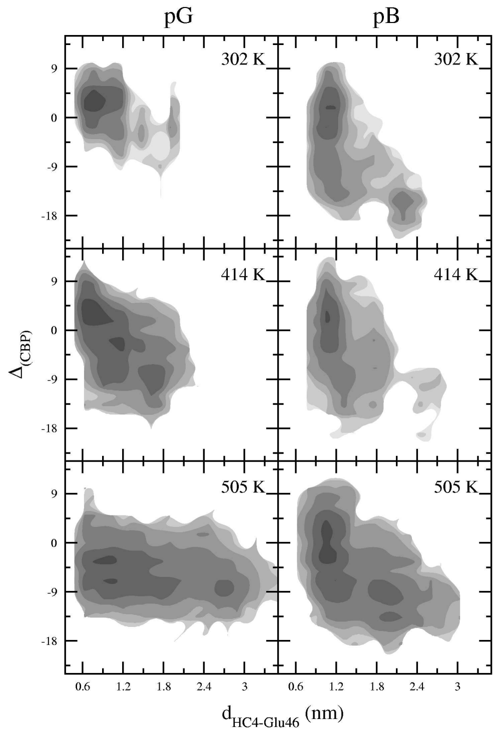
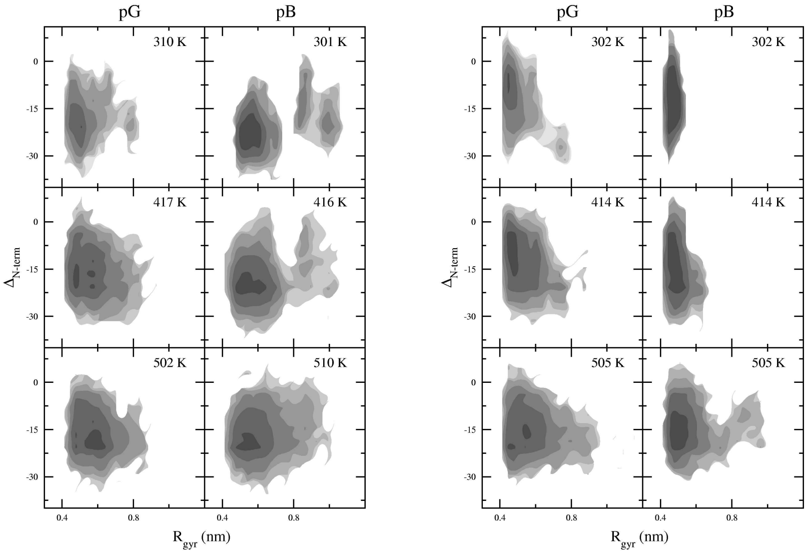


Figure 3.5: **Free energy diagrams at three different temperatures** as a function of the distance between the chromophore and Glu46, and the hydrogen bond difference Δ_{CBP} (Eq. 3.1) in the chromophore binding pocket for PT xtal simulations of the pG and pB state, indicated by the labels. The contour lines indicate the $k_B T$ levels, decreasing with darker shading. White areas have not been sampled.

$\Delta_{\mathrm{N-term}}$ has more negative values in comparison to Δ_{CBP}, ranging from $\Delta_{\mathrm{N-term}}$ = 11 for the crystal structure to $\Delta_{\mathrm{N-term}}$ = -40 for complete solvation. Fig. 3.6a shows the free energy profiles for the N-terminal domain for the nmr-based pG and pB PT simulations, while Fig. 3.6b shows those for the crystal based simulations. The conformations sampled for the N-terminal cap at around 300 K in the pG simulations have a minimum at (R_{gyr} = 0.50 nm, $\Delta_{\mathrm{N-term}}$ = -16) for the nmr simulation and at (R_{gyr} = 0.45 nm, $\Delta_{\mathrm{N-term}}$ = -5) for the crystal simulation. The values for $\Delta_{\mathrm{N-term}}$ differ significantly, relating to the fluctuating N-terminal helices in the nmr simulation and the well-defined helical structures in the xtal simulation. The free energy profile of the pG crystal simulation shows a second minimum at R_{gyr} = 0.75 nm and $\Delta_{\mathrm{N-term}}$ = -25, closely resembling a similar state in the nmr-pG simulation. The large values for the radius of gyration in this state indicate that expansion of the hydrophobic core is closely linked to loss of α-helical structure. At higher temperatures, the free energy profiles of both pG simulations are similar.

The free energy profiles for the nmr and crystal simulations exhibit large differences, see Fig. 3.6. In the xtal simulation of the pB state the hydrophobic core is very compact, since R_{gyr} expands beyond 0.6 nm only at a temperature of 505 K. $\Delta_{\mathrm{N-term}}$ varies widely, from 5 to -30, unrelated to the core compactness. In contrast, the pB-nmr simulation free energy profile has



(a) nmr (b) crystal

Figure 3.6: **Free energy profiles of the N-terminal cap** in the receptor state pG and the signaling state pB as a function of the radius of gyration of the hydrophobic core and the hydrogen bond difference $\Delta_{\mathrm{N-term}}$ (Eq. 3.1) in the helical residues, with (a) nmr simulations and (b) xtal simulations. The contour lines indicate the k_BT levels, decreasing with darker shading. White areas have not been sampled.

an minimum at R_{gyr} = 0.60 nm with few intra-protein hydrogen bonds, since $\Delta_{\mathrm{N-term}}$ varies between -18 and -27. Separated by a barrier, a second minimum at R_{gyr} = 0.80 nm and less negative values for $\Delta_{\mathrm{N-term}}$ occurs that also appears in the pG simulations. A third state, at R_{gyr} = 1.00 nm, $\Delta_{\mathrm{N-term}}$ = -20) represents a widely expanded hydrophobic core with little α-helical hydrogen bonds. At higher temperatures the barrier separating these states decreases to a few k_BT and eventually disappears at 510 K.

Discussion

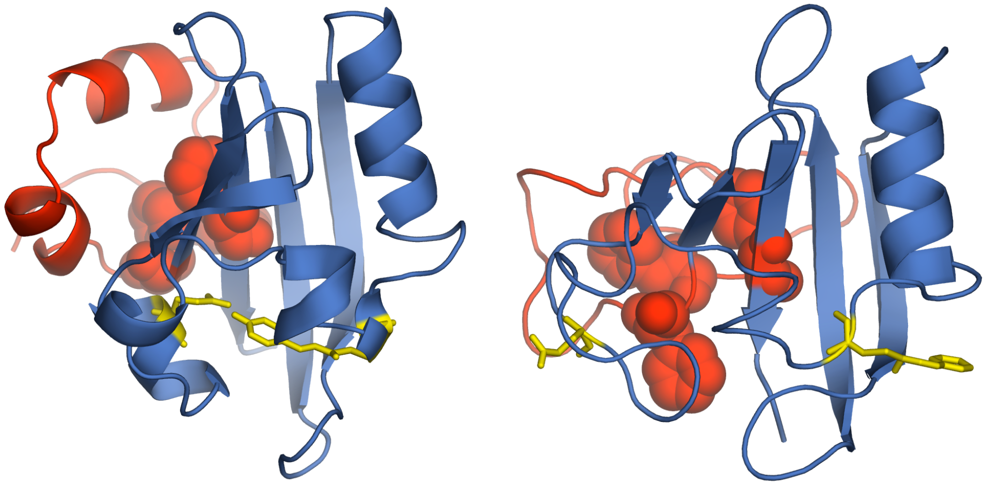
The main interest in this work is the elucidation of the conformational transitions during the photo-cycle of the Photoactive Yellow Protein (PYP), especially those linked to the formation of the signaling state. Several experiments hinted at the extent and details of the structural changes that occur during pB formation, including molecular simulation studies [75, 77, 78], although, it is inefficient to use MD simulations to access the long timescales necessary for escaping a local minimum. This has prevented most studies from sampling the signaling state. Parallel Tempering (PT) combines conventional MD replicas at different temperature with a Monte Carlo scheme for exchanging temperatures. In this way PT overcomes free energy barriers, and hence enabled us to expand the exploration of the configurational space of PYP. The implementation of the PT algorithm in the case of PYP involved the choice which starting point would be most relevant. Complete and well-defined crystal structures are most accurate and thus serve best to initiate molecular simulation procedures, whereas solution structures suffer from the fact that less well-defined protein regions introduce additional flexibility [99]. On the other hand, Rajagopal *et al.* state that the crystalline lattice restrains PYP, and the N-terminal domain in particular, from losing helical structure during the signaling process [100].

In our PT simulations, initial starting structures do not result in different time-averaged fluctuation profiles, as illustrated in Fig. 3.2. The average fluctuation per residue does not differ significantly for simulations based on coordinates from X-ray diffraction experiments or NMR spectroscopy. Looking in more detail at the chromophore binding pocket (Fig. 3.4 and Fig. 3.5), a similar picture emerges: loss of structure, exposure of protein interior and the intrusion of water molecules in the protein core occur at a similar level in simulations that were initiated from different starting structures. However, there is an exception, related to a topic of debate in the literature: the role and extent of unfolding of the N-terminal cap. In the crystal simulations the N-terminal domain is more compact, with clearly defined helices, whereas in the nmr simulations it shows a more loose conformation in which water molecules can enter more easily. The latter conformation agrees with the observation that the interaction energy between water molecules and the protein is higher in the nmr simulations than for the xtal simulations. At higher temperatures, the crystal simulations also visit the more loose conformation. The structure of the N-terminal domain in crystal structures may represent a conformation induced by crystal contacts that is packed too tightly to represent the protein in an aqueous environment, in agreement with the crystallographic work on a mutant, E46Q, of PYP [100, 101].

The fluctuations shown in Fig. 3.2 indicate that the second N-terminal helix is less well defined in comparison to the first. Residues 11-15 have in each PT simulation lower average fluctuations than residues 19-23. Imamoto *et al.* find that the removal of the second helix does not affect the change in radius of gyration of the protein when exposed to light, whereas the removal of the first helix induces an increase in volume during the photo-cycle. Moreover, removing the first helix also affects the structural change occurring in the central β-sheet [67]. These observations agree well with the explanation that the second α-helix in the N-terminal domain fluctuates between two configurations, a well-structured, helical form and a disordered, loop-like form. Both occur in our simulations of the receptor and the signaling state, although in the latter only at higher temperatures for the crystal simulations. The first α-helix in the N-terminal cap is ordered in the pG state, but loses structure, and suffers water intrusion upon solvent exposure of the chromophore.

Fig. 3.4 and Fig. 3.5 lead to the conclusion that a different chemical composition (pG vs. pB' or pB) of the chromophore binding pocket leads to a different free energy profile. If the temperature is sufficiently high, water molecules enter the binding pocket and disrupt its integrity regardless of its state. The location of the negative charge determines the mechanism of CBP-disruption. In the receptor state pG, the chromophore contains a negative charge, delocalized over its whole length. A strong interaction exists between the chromophore and the positively charged Arg52, the latter being in contact with solvent molecules. A fluctuation causing Arg52 to move more toward the solvent leaves the chromophore prone to solvent exposure. Balancing the favorable interaction with solvent are surrounding residues that contribute to a hydrogen bonding network that further stabilizes the negative charge on the chromophore. When these connections inside the CBP are broken at high temperature, the chromophore shifts toward the solvent and disrupts the chromophore binding pocket.

In the case of the signaling state pB, the negative charge is located at Glu46 and localized over a smaller set of atoms in comparison to the chromophore. This has two consequences with regard to the CBP-disruption mechanism: first, the negatively charged Glu46 destabilizes the hydrogen bonding network inside the protein and allows water molecules to access the protein interior and second, the interaction between Arg52 and the chromophore has become less favorable. Consequently, the sequence of events has reversed in the signaling state with respect to the receptor state, first Glu46 becomes solvent exposed followed by emergence of the chromophore into the solvent. The final situation in both the pG and pB states is the same, as depicted in Fig. 3.7: a huge disruption of the chromophore binding pocket, resulting in loss of α-helical content, in agreement with literature. Of course, the most important difference is that at room temperature the CBP disruption in receptor state pG is much more unlikely than in the pB state, as is found in experiments.


(a) pG

(b) pB

Figure 3.7: **Ribbon representation of PYP** in (a) the crystal structure of the receptor state pG and (b) a typical conformation of the signaling state pB at ambient conditions, taken from the PT run and clearly exhibiting pB features. Red indicates the N-terminal cap, with the Phe6, Phe28 and Phe121 in space filling model to represent the hydrophobic core. The yellow stick models represent the chromophore and Glu46.

The conformational rearrangements in the CBP relate to those in the N-terminal cap. When Glu46 becomes fully solvent exposed, it interacts with the backbone of the N-terminal hydrophobic core, thus disrupting it. This observation leads to the conclusion that our simulations have predicted the conformation of the signaling state pB of PYP. This state has the following characteristics: The hydrogen bonding network in the CBP has disappeared, and both the chromophore and Glu46 are fully solvent exposed. These rearrangements cause the loss of α-helical structure in the first and last helix in the PAS-core. Glu46 interacts with the N-terminal cap, causing conformational instabilities in the N-terminal hydrophobic core and α-helices. Fig. 3.7

summarizes these observations and shows the crystal structure of the receptor state next to a typical room temperature conformation of the signaling state. Recent results from NMR experiments on a truncated form of PYP, where removal of the N-terminal cap has led to a extended pB life time, agree very well with our prediction [102].

In this work, chromophore protonation occurred through removal of the proton at Glu46 to place it at the chromophore, neglecting energetic considerations, such as whether the protein environment had assumed a configuration favorable for proton transfer. Although a mechanism involving a direct proton transfer mechanism from Glu46 to the chromophore is certainly possible, this manual transfer is probably too crude to describe the change of protonation states in the chromophore binding pocket. Our PT results indicate that the proton transfer in PYP during its photo-cycle might actually be more complicated, involving solvent intermediates. The reverse reaction, relevant for the recovery of the receptor state, also may include multiple pathways.

We should stress that the parallel tempering simulations, although expanding the exploration of the PYP conformation space, are still not completely converged. The complete equilibration of all states requires that each replica makes many trips from the lowest to the highest temperature, which might take a multiple amount of the simulation time yet invested. Another caveat is that, although the room temperature simulations are at ambient pressure, the high temperature simulations are at unphysical high pressures and cannot be used to compare with experiment. Also, the (GROMACS) force field might have deficiencies and underestimate the stability of partially unfolded protein structures. An exhaustive comparison between different forcefields is beyond the scope of the present work. However, despite all this we believe that the qualitative results obtained in this work are reproducible and that the main conclusions are warranted.

Concluding remarks

In this work we have used parallel tempering simulations to study the conformational transitions associated with the photo-cycle of the Photoactive Yellow Protein. The main goal of this work is the prediction of the mechanism of formation and the structure of the signaling state pB (see Fig. 3.7). Comparing several independent PT simulation series we found the following mechanism for pB formation. After the initial isomerization and proton transfer from Glu46 to the chromophore, the negative charge is located at Glu46 and localized over a smaller set of atoms in comparison to the previous location of the negative charge at the chromophore. This negative charge has two effects on the chromophore binding pocket: first, the negatively charged Glu46 destabilizes the hydrogen bonding network inside the protein and allows water molecules to access the protein interior and second, the interaction between Arg52 and the chromophore has become less favorable. The next step is the solvent exposure of Glu46, followed by the solvent exposure of the chromophore. The solvent exposure of Glu46 has also an effect on the stability of the hydrophobic core in the N-terminal domain, resulting in the partial unfolding seen in several experiments. In summary, we have predicted the structure and the formation mechanism of the signaling state in the photo-cycle of PYP. Our results are qualitative, but compare well to very recent NMR experiments [102].

The PT has proved very powerful in sampling rugged energy landscapes such as occur in protein conformational transitions. However, the technique cannot give detailed information of the kinetics of the conformation transitions. In the near future we will employ other advance simulation techniques such as transition path sampling and related techniques to access the relevant kinetic information in PYP [103].

Chapter 4

Conformational requirements for an efficient recovery reaction of Photoactive Yellow Protein[*]

Previous simulations have shown that the formation of the signaling state of PYP is characterized by the solvent exposure of the chromophore and Glu46 [104]. These results are in agreement with NMR spectroscopy data obtained from Δ_{25}*-PYP [102]. In this work we have compared our conformational prediction of the signaling state with the structural ensemble obtained with NMR. A parallel tempering simulation of the truncated mutant enables direct comparison with experiment, allowing further validation of our previous prediction of the signaling state. Furthermore, this comparison gives insights into the role of the N-terminal domain. Also, we have attempted to sample the recovery of the receptor state i.e. the conformational requirements for the protein to return to its receptor state. Initially, we focused on the conformational characteristics of the receptor state in more detail. Using different conformations along the recovery reaction pathway we present parallel tempering simulations that highlight the conformational aspects of the recovery of the receptor state.*

Introduction

Photoactive Yellow Protein (PYP) is a water soluble blue-light photoreceptor from *Halorhodospira halophila* [25, 50]. Comprising 125 amino acids and a covalently bound chromophore (trans-4-hydroxy cinnamic acid) [51], the protein folds into a PAS core capped by an N-terminal domain containing two helices [20, 26]. Upon absorbing a blue-light photon as a trigger, PYP undergoes a photo cycle, starting from its receptor state (pG). Visiting several intermediate states, the chromophore twists along its double bond to a *cis* configuration within picoseconds [29]. Within milliseconds of the isomerisation, a proton from Glu46 (protonated in the receptor state) *trans*fers to the chromophore, leaving a negative charge on Glu46 [53, 54]. Driven by the new negative charge in the chromophore binding pocket, the protein unfolds to expose the chromophore and Glu46 to bulk water [102, 104, 105], forming the signaling state (pB), blue-shifted with respect to the receptor state [106]. Completion of the cycle, *i.e.* the refolding of the protein to the receptor

---

[*] In preparation

state, requires several hundreds of milliseconds.

As PYP is an easily accessible system for a wide variety of techniques, many of its characteristics have been elucidated. For example, the initial events of the photo cycle have become much clearer recently with the use of ultrafast spectroscopy techniques [107], time resolved X-ray crystallography [108] and molecular simulation studies [73]. However, a few issues remain unclear, such as the role of the N-terminal domain during the photo cycle. The mutant Δ_{25}-PYP proves that this part of the protein is involved in the photo cycle: This mutant lacks the N-terminal cap, via truncation of the first 25 amino acids. As a result, the recovery of the pG state from the blue-shifted signaling state pB is significantly retarded [35], although the crystal structure of the receptor state does not show large differences [109] with respect to its wild-type counterpart.

In contrast to the wealth of data on the initial photo cycle events, the slower process of the recovery of the receptor state is still in need of a more detailed characterization. Reformation of pG requires several chemical and conformational rearrangements: *cis* to *trans* isomerisation of the chromophore, deprotonation of the chromophore, protonation of Glu46 and refolding of the protein. In the pB state the chromophore is in the *cis* configuration and thermally reisomerises back to its pG *trans* configuration. Upon absorption of another photon this isomerisation process occurs instantaneously and speeds up the recovery process by a 1000-fold. This is known as the branching reaction and revealed the isomerisation reaction as the rate limiting step [110]. Before the re-isomerisation occurs, the chromophore loses a proton at the phenolate oxygen, as evidenced by the observation of the intermediate pB^{deprot} [55]. Since the formation of the negative charge on Glu46 is the driving force of the formation of the signaling state (*i.e.* the partial unfolding of the protein), it is clear that refolding is unlikely to occur with a negative charge located on this residue. The solvent most probably provides the proton for reprotonation of Glu46 [55].

Previous simulations have shown that the formation of the signaling state of PYP is characterized by the solvent exposure of the chromophore and Glu46 [104]. These results are in agreement with NMR spectroscopy data obtained from Δ_{25}-PYP [102]. In this work we have compared our conformational prediction of the signaling state with the structural ensemble obtained with NMR. A parallel tempering simulation of the truncated mutant enables direct comparison with experiment, allowing further validation of our previous prediction of the signaling state. Furthermore, this comparison gives insights into the role of the N-terminal domain. Also, we have attempted to sample the recovery of the receptor state *i.e.* the conformational requirements for the protein to return to its receptor state. Initially, we focused on the conformational characteristics of the receptor state in more detail. Finally, using different conformations along the recovery reaction pathway we present parallel tempering simulations that highlight the conformational aspects of the recovery of the receptor state.

Methods

In this work we have used parallel tempering simulations to (i) investigate the role of the N-terminal domain in the photo cycle of PYP, (ii) probe the conformational characteristics of the receptor state and (iii) attempt to sample the refolding event associated with the recovery of the receptor state.

Systems

For the comparison of the conformation of the signaling state obtained with parallel tempering simulations and NMR spectroscopy, we performed two parallel tempering simulations. Both started from the NMR structure of wild type PYP in its receptor state (PDB-code 3PHY, conformation 11), with manual alterations to reflect the different chemical topology of the chromophore binding pocket (*i.e.* protonated *cis* chromophore and deprotonated Glu46). The difference between the two simulations is that one contains the full-length protein, denoted as pB′, while the other lacks the first 25 N-terminal residues, denoted as Δ_{25}-pB′. In addition, we performed a parallel tempering simulation of the receptor state, denoted as pG, initiated from the conformation used as starting structure for the pB′ simulation

We selected a conformation from our previously published pB simulation [104] as starting point for three simulations: pB^{deprot}, $\text{pB}^{deprot}_{flexible}$ and pG′, aiming to sample different aspects of the recovery trajectory of PYP. Table 4.1 lists the details of the topologies of the systems. In all three simulations the protonation states resemble the receptor state: a deprotonated chromophore and protonated Glu46. The orientation of the chromophore differs: pB^{deprot} contains a *cis*-chromophore and pG′ contains a *trans* chromophore. The $\text{pB}^{deprot}_{flexible}$ simulation starts with a *cis*-oriented chromophore, but has a lowered rotation barrier for the double bond, allowing isomerisation to occur within the duration of the simulation. To lower the rotation barrier around the double bond, we reduced the dihedral potentials around the double bond with approximately one-third. This resulted in *cis* to *trans* isomerisation at temperatures above 475 K.

System	Starting structure	Isomer	Chromophore	Glu46	Runtime (ns)
pG	NMR of pG	*trans*	deprotonated	protonated	10
pB′	NMR of pG	*cis*	protonated	deprotonated	10
Δ_{25}-pB′	NMR of pG	*cis*	protonated	deprotonated	10
pB^{deprot}	simulation of pB	*cis*	deprotonated	protonated	10
pG′	simulation of pB	*trans*	deprotonated	protonated	15
$\text{pB}^{deprot}_{flexible}$	simulation of pB	*cis*	deprotonated	protonated	10.5

Table 4.1: **System details.** The labels *cis* and *trans* indicate the configuration of the chromophore, whereas the labels protonated and deprotonated indicate the protonation states of the chromophore and Glu46 respectively. The runtime is the simulation time per replica. All parallel tempering simulations contained 64 replicas.

System preparation

Each system listed above underwent an equilibration procedure, before starting parallel tempering. First, water and hydrogen positions are optimized during a 20 ps MD run with the heavy protein atoms restrained. Second, 1 ns of MD provided equilibration of the complete system in the NpT ensemble, using the Berendsen thermostat [97] (τ_T = 0.1 ps) and barostat (τ_p = 1.0 ps). Addition of water and ions has been described elsewhere [104].

Parallel tempering

We performed parallel tempering runs for the systems described above using a Perl script in combination with the Gromacs software package. Each PT run contained 64 replicas undergoing exchange attempts every 1 ps. The 1 ps MD runs were performed using GROMACS [91], with a time step of 2 fs. The Berendsen thermostat (τ_T = 0.1 ps) imposed a temperature on the replicas. Coupling to the thermostat of the protein and the solvent, including the ions, was separate. The temperature in the parallel tempering runs varied between 282 K and 645 K. The temperature gap was initially estimated by a linear dependence on the inverse temperature, and turned out afterward to give rise to a uniform acceptance ratio of around 20%. A Perl script governed the swaps of temperatures between the replicas. Every picosecond 4032 swap attempts were performed.

The acceptance rule to exchange two replicas at temperatures i and j is:

$$P_{acc}(ij) = min\left(1, e^{\Delta\beta_{ij}\Delta U_{ij}}\right) \tag{4.1}$$

with $\Delta\beta_{ij}$ the difference of the inverse of the swapping temperatures and ΔU_{ij} the energy difference of the two configurations. The temperature distribution over the replicas resulted in an average acceptance ratio of 20 %.

Figure 4.1 shows the temperature trajectories for each replica, coloured according to their starting temperature.

After 2 ns some of the low temperature replicas have reached high temperatures and vice versa (Fig. 4.1). With an average acceptance ratio of 20 % the replicas do not visit both high temperatures and low temperatures within the simulation time of 10 ns.

Analysis

Conformations at similar imposed temperatures were combined into trajectories, excluding the first 2 ns. Analysis of these trajectories consisted of tracking several order parameters using the tools included in the Gromacs software package in combination with perl scripts. The order parameters used here to investigate the conformational barriers in the system are the hydrogen bond difference [111] of the residues in the chromophore binding pocket (Tyr42, Glu46, Thr50, Arg52, Cys69 and chromophore, Phe96, Tyr98, Met100) and the distance between the centers of mass of several groups in the chromophore binding pocket. Also, we followed the number of water molecules within a radius of 3.5 Å around the chromophore and around Glu46, and the number of hydrogen bonds between the backbone atoms of residues 42 to 59, comprising two helices. Finally, the dihedral angle along the double bond of the chromophore was sampled.

Sampling of rejected states

The parallel tempering simulations presented here are not fully equilibrated. Although high temperature replicas visit low temperatures and vice versa, not all replicas visit all temperatures within the simulation time of 10 ns. System size is prohibitive, and as such, Photoactive Yellow Protein is close to the limit of the parallel tempering scheme in combination with current CPU power. As a consequence, free energy differences obtained from these simulations are

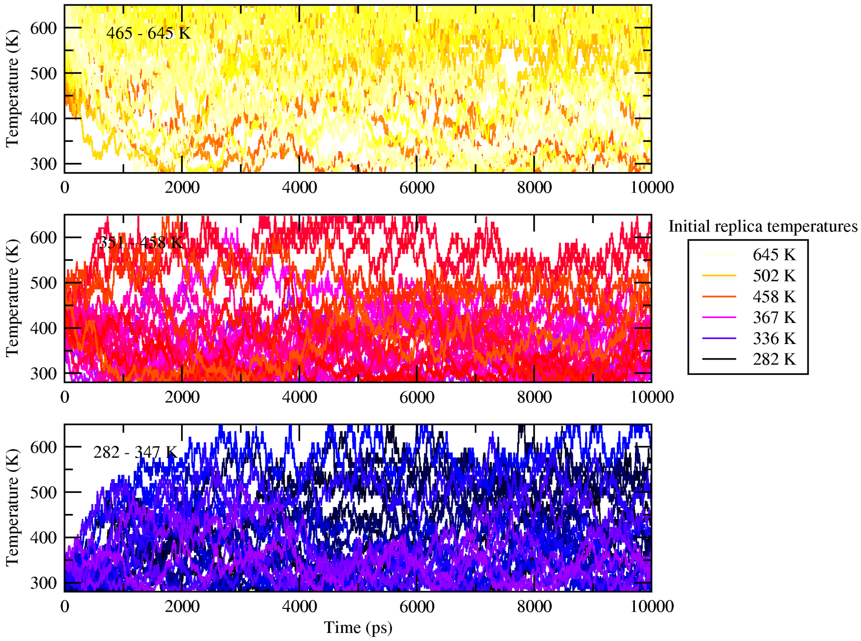


Figure 4.1: **Imposed replica temperature as a function of time.** The colour of each replica indicates its starting temperature, starting at black (T = 282), going to blue and red, and ending at yellow (T = 645 K). For clarity, replicas 0-21 (lower), 22-43 (middle) and 44-63 (upper) are depicted in separate graphs. The graphs shows that high temperature replicas reach low temperatures and vice versa.

inaccurate. The free energy difference in replica i, as function of an order parameter Q is:

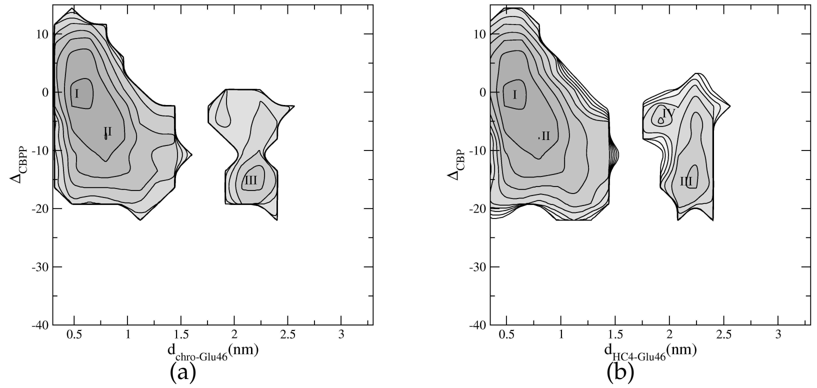
$$F_i = -k_B T_i \ln P_i(Q) \tag{4.2}$$

with k_B the Boltzmann factor, T the temperature and $P_i(Q)$ the probability distribution. $P(Q)$ is a histogram of the occurrence of order parameter Q.

A recent paper by Frenkel [112] states that Monte Carlo schemes may improve significantly in accuracy by including the properties of the rejected states. Another paper applies the sampling of rejected states to a parallel tempering scheme [113]. Effectively, this means that the contributions of different replicas to the histogram of order parameter Q are scaled with their Boltzmann factors. The probability distribution can then be estimated as:

$$P_i(Q) = \sum_{j=1, j\neq i}^{N} \left(1 - e^{\Delta\beta_{ij}\Delta U_{ij}}\right) \delta(Q_i - Q) + \sum_{i=1, i\neq j}^{N} \left(min\left(1, e^{\Delta\beta_{ij}\Delta U_{ij}}\right)\right) \delta(Q_j - Q) \tag{4.3}$$

with i and j replica indices and N the total number of replicas.



(a) (b)

Figure 4.2: **Free energy profiles of the formation of the signaling state.** The profiles are calculated as a function of the distance between the chromophore (HC4) and Glu46, and the hydrogen bond difference Δ_{CBP} (Eq. 3.1) in the chromophore binding pocket (CBP). (a) Order parameters are calculated for conformations at 301 K in the pB′ simulation (b) Rejected states also contributed to the free energy profile at 301 K of the pB′ simulation. (c) Order parameters are calculated for conformations at 301 K in the Δ_{25}-pB′ simulation. The contour lines indicate the k_BT levels, decreasing with darker shading. White areas have not been sampled. The labels I-IV indicate free energy wells, with I at ($d_{HC4-Glu46}$ = 0.6 nm, Δ_{CBP} = 0), II at ($d_{HC4-Glu46}$ = 0.8 nm, Δ_{CBP} = -7), III at ($d_{HC4-Glu46}$ = 2.2 nm, Δ_{CBP} = -15) and IV at ($d_{HC4-Glu46}$ = 1.9 nm, Δ_{CBP} = -5).

Results and discussion

In this work we present parallel tempering simulations to (i) provide a validation of our previously published prediction of the signaling state conformation of Photoactive Yellow Protein (PYP) [104], (ii) further investigate the conformational characteristics of the receptor state pG of PYP, and (iii) characterize the conformational aspects involved in the recovery reaction of the PYP photo cycle.

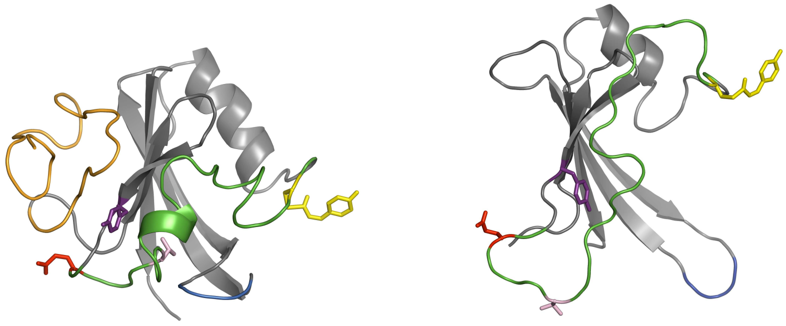
Including the contribution of the rejected states

As discussed in the Methods section, the parallel tempering simulations presented here are not fully equilibrated. As a consequence, free energy differences obtained from these simulations are inaccurate. Including the rejected states in the calculation of the free energy, accuracy can improve [112,113]. Using the pB′ simulation, we calculated two free energy profiles of the chromophore binding pocket at 301 K. Figure 4.2(a) shows the free energy profile calculated from conformations at 301 K and figure 4.2(b) shows the free energy profile where the rejected states also contribute. The minima in the free energy profile at 301 K are indicated with labels I, II and III and respectively represent conformations with an intact chromophore binding pocket, including the hydrogen bond between the chromophore and Glu46; conformations in which Glu46 is no longer part of the hydrogen bonding network around the chromophore; and conformations in which also the chromophore has become exposed to solvent. Note that the distances are calculated between centers of mass, and do not indicate distances between hydrogen bond donors and acceptors. A barrier of 1 k_BT separates I and II, whereas the barrier between II and III is higher than 7 k_BT. Including the contribution of the rejected states in the calculation of the free energy profile results in a similar shape of the profile. The three minima I, II and III appear as well as an additional minimum, labeled IV. This minimum represents conformations where the chromophore is exposed to solvent, but Glu46 still interacts with residues in the chromophore binding pocket. Including the rejected states in the calculation of the profile results in a smoother profile. All other free energy profiles presented here include the contribution of the rejected states.

The role of the N-terminal domain

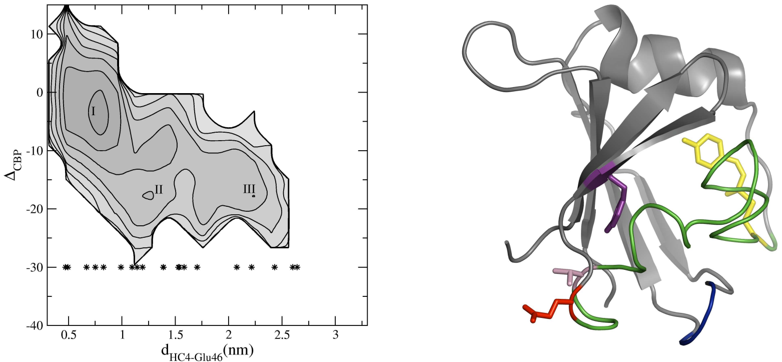
The mutant Δ_{25}-PYP lacks the N-terminal cap, and as a result, the recovery of the pG state from the blue-shifted signaling state pB is significantly retarded [35]. This pB lifetime is sufficiently long to enable structure determination with NMR spectroscopy [102]. Figure 4.3 shows a typical conformation of the signaling state obtained with parallel tempering simulations [104] and the pB NMR structure (PDB-code 1XFQ), with the central β-sheet perpendicular to the plane of the paper. Both conformations have an intact central β-sheet with Glu46 and the chromophore oriented towards the solvent. The region containing these two groups has a different shape in the two conformations: the simulation structure has a higher α-helical content. Another difference between the two conformations is that the NMR conformation is more extended. This is partly due to the higher degree of unfolding, but also due to a different orientation of the loop containing Met100 (highlighted in blue). In the simulation, this loop is oriented toward the Glu46-region, whereas it points outward in the NMR-conformation. We conclude that the presence of the N-terminal cap clearly influences the degree of unfolding, upon signaling state formation. Comparison of the two side views (figures 4.3(a) and 4.3(b)) reveals that the N-terminal cap is close to the region containing Glu46, thereby restricting the conformational freedom of Glu46 during its solvent exposure process.

To understand the differences observed in snapshots 4.3(a) and 4.3(b) we performed a simulation of pB′, with the first 25 amino acids deleted. Figure 4.4(a) shows the free energy profile as a function of the distance between the chromophore and Glu46 and the hydrogen bond difference. The deepest well in the free energy profile is indicated as I and represents conformations in which both groups are still within the protein core, but no longer hydrogen bonding. The overall shapes of the free energy profiles of pB′ (figure 4.2(b) and Δ_{25}-pB′ are clearly different. A more detailed comparison between the two free energy profiles shows that minimum III is present in the Δ_{25}-pB′ simulation as a shallow well; that the minimum indicated as IV in figure 4.2(b) is absent; and that the Δ_{25}-PYP simulation has an additional minimum, denoted as II in figure 4.4(a). The latter minimum represents conformations with Glu46 solvent exposed and the chromophore located within the protein. Most striking, however, is the absence of a barrier between I and III. These observations indicate that Δ_{25}-PYP is less restricted in its conformational space and that the barrier for unfolding is significantly lowered.



(a) PT (side view) (b) NMR (side view)

Figure 4.3: **Comparison of signaling state conformations obtained from simulation and experiment.** (a) shows a snapshot from the pB simulation in ref. [104]. (b) shows conformation 7 from the NMR structure of the Δ_{25}-PYP pB state. The ribbon representation shows the protein backbone in grey, with residues 45-70 highlighted in green, residues 99-101 highlighted in blue and residues 1-25 highlighted in orange, if present. The stick models display the side chains of residues Glu46 (red), Thr50 (pink) and Cys69-HC4 (yellow). Hydrogen atoms are not shown.



(a) (b)

Figure 4.4: **Parallel tempering simulation of Δ_{25}-PYP-pB′** (a) Free energy profile as a function of the distance between the chromophore (HC4) and Glu46, and the hydrogen bond difference Δ_{CBP} (Eq. 3.1) in the chromophore binding pocket (CBP). The contour lines indicate the k_BT levels, decreasing with darker shading. White areas have not been sampled. The labels I-III indicate free energy wells, with I at ($d_{HC4-Glu46}$ = 0.8 nm, Δ_{CBP} = -3), II at ($d_{HC4-Glu46}$ = 1.3 nm, Δ_{CBP} = -18) and III at ($d_{HC4-Glu46}$ = 2.2 nm, Δ_{CBP} = -18). The stars indicate the distance between HC4 and Glu46 in the NMR structure of Δ_{25}-PYP-pB. Their position at Δ_{CBP} = -30 has no meaning. (b) Snapshot from the $\Delta 25$-pB′ simulation. The ribbon representation shows the protein backbone in grey, with residues 45-70 highlighted in green, residues 99-101 highlighted in blue and residues 1-25 highlighted in orange, if present. The stick models display the side chains of residues Glu46 (red), Thr50 (pink) and Cys69-HC4 (yellow). Hydrogen atoms are not shown.

To compare the Δ_{25}-PYP-pB′ simulation to the NMR structure of the pB state of this mutant, figure 4.4(a) shows the distance between the chromophore and Glu46 as stars (with an arbitrarily chosen value for Δ_{CBP}) for the NMR structural ensemble. The HC4-Glu46 distance in the NMR structure extends beyond 2.5 nm, a value hardly sampled in the Δ_{25}-PYP-pB′ simulation and even less sampled in the wild type pB′ simulation. More interesting is the occurrence of stars in the region between 1.5 and 2 nm, the place where a high barrier separates the conformations with an intact CBP from the partially unfolded conformations in the wild type pB′ simulation. This would suggest that the NMR structural ensemble would compare better to the pB′ simulation of Δ_{25} simulation than to the wild type.

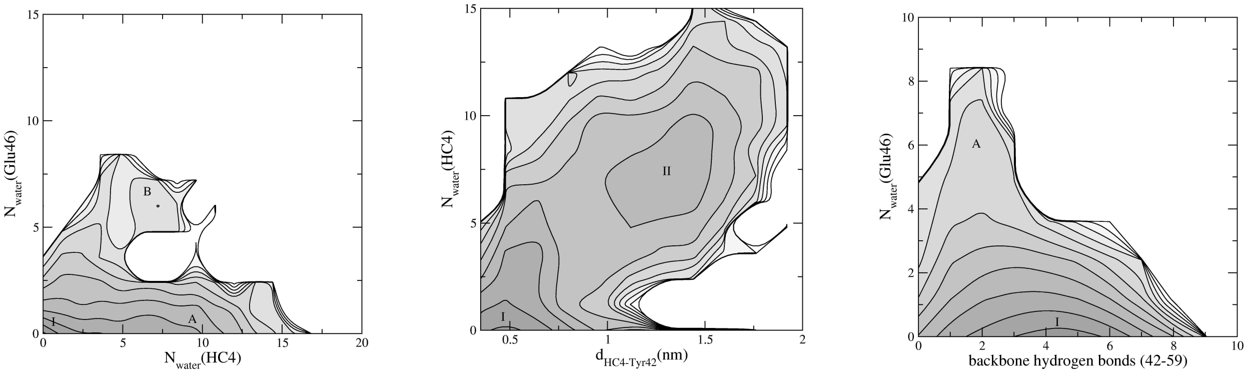
Figure 4.4(b) shows a typical snapshot from the Δ_{25}-pB′. This snapshot shows that the central β-sheet has an extended and twisted conformation. The chromophore is oriented toward the protein core, but not buried in it. The flexibilities in the β-sheet facilitate more orientations for the large helix flanking the β-sheet, as well as for the region containing Glu46. As a consequence, the chromophore is more accessible for water molecules. The combination of the enhanced flexibility of the β-sheet and the absence of the N-terminal cap enable Glu46 to move away further from the protein core. The snapshot of the NMR structure, figure 4.3 (b), shows a similar extended conformation for the β-sheet, but lacks the twisted orientation. Also, the chromophore is oriented toward solvent, as well as Glu46. Since the data in this region, highlighted green in figure 4.3, is limited [102], many conformations are compatible with the available structural constraints, including a more compact one.

Comparing the three snapshots in figures 4.3 and 4.4(b) shows that removal of the N-terminal cap leads to increased flexibility in the central β-sheet and in the region containing Glu46. Previously, we have shown that the helical structure in the N-terminal domain is lost during the formation of pB [104]. Also, the hydrophobic core expands, but does not completely unfold. In Δ_{25}-PYP, part of the hydrophobic core (Phe121 and Trp117) is exposed to solvent, which destabilizes the protein and leads to extended unfolding of Δ_{25}-PYP in the signaling state. The N-terminal cap shields the hydrophobic convex surface of the β-sheet of the protein, and thereby restricts the conformational freedom of the protein.

We show that the signaling states of wild type pB and Δ_{25}-PYP exhibit differences, all related to the N-terminal cap restricting the conformational freedom of the protein. Observations on the recovery kinetics of both wild type and the truncated mutant protein show that the recovery of the receptor state occur faster in the wild type [35], too fast for full structure determination with NMR [71]. An even faster recovery reaction occurs in the E46Q mutant, in which less unfolding takes place, due to the lack of a negative charge on the position of Glu46 (Gln46 in the mutant) [72]. With the comparison between simulation and NMR spectroscopy of the signaling state of Δ_{25}-PYP, we have validated our prediction of the signaling state of wild type PYP.

Conformational characteristics of the receptor state pG

Our previous simulations of the receptor state of PYP, pG, sampled conformations with the chromophore interacting with solvent molecules, outside the chromophore binding pocket. Also, Glu46 became solvent exposed. These two alterations resulted in an increase of the distance between the chromophore and Glu46 and a disruption of the hydrogen bonding network in the chromophore binding pocket. Although we already performed this simulation earlier [104], we



(a) (b) (c)

Figure 4.5: **Free energy profiles of the receptor state simulation.** These profiles are calculated as a function of (a) the number of water molecules within a radius of 0.35 nm around the chromophore and around Glu46; (b) the distance between the chromophore and Tyr42, and the number of water molecules within a radius of 0.35 nm around the chromophore; and (c) the number of backbone hydrogen bonds of residues 42-59 and the number of water molecules within a radius of 3.5 Å around Glu46. The contour lines indicate the k_BT levels, decreasing with darker shading. White areas have not been sampled. The roman labels indicate free energy wells, the labels A and B indicate directions.

repeat it here to gain more insight into the solvent structure in and around the chromophore binding pocket.

Figure 4.5(a) shows the free energy profile of the receptor state simulation as a function of the number of water molecules within a radius of 0.35 nm around the chromophore and Glu46. The deepest well in the profile, labeled I, represents conformations where both the chromophore and Glu46 have no contact with water molecules: The chromophore binding pocket (CBP) is intact. From this minimum the profile extends into two directions: A, with little solvent exposure for Glu46, and B, with Glu46 interacting with up to eight water molecules, labeled B. The chromophore is in contact with solvent in both directions. This figure clearly shows that solvent exposure of the chromophore occurs, as well as solvent exposure of Glu46, and that the latter occurs less often than the first. Since the parallel tempering simulations of PYP are not well equilibrated (*i.e.* not all replicas have visited all temperatures), it is not possible to be more specific.

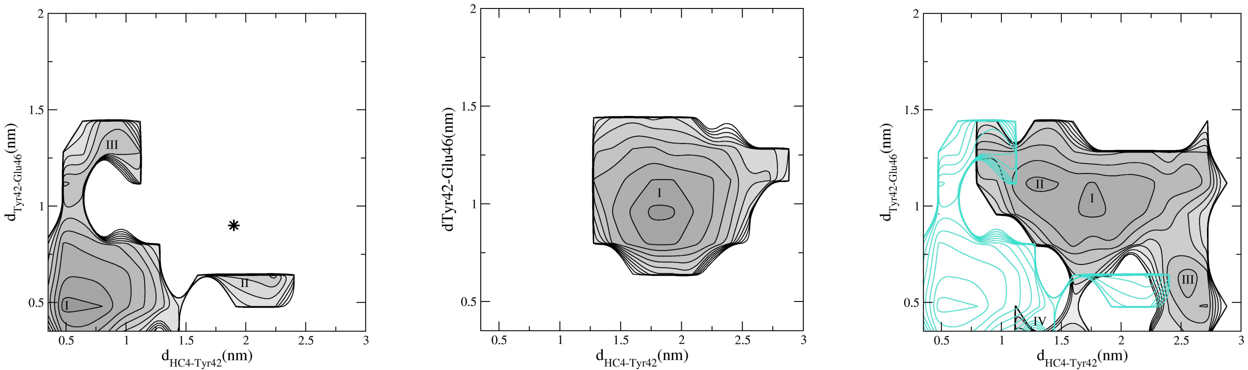
In an intact CBP a hydrogen bond exists between Tyr42 and the chromophore. Fig. 4.5(b) displays the free energy profile using the distance between the centers of mass of the chromophore phenolate ring and the side chain of Tyr42, and the number of water molecules within a radius of 0.35 nm surrounding the chromophore. With no water molecules around the chromophore, there is a minimum at $d_{HC4-Tyr42}$ = 0.5 nm, (labeled I), implying a hydrogen bond between the chromophore and Tyr42. A barrier of 4 k_BT separates this well from minimum II, representing more loose protein conformations, including conformations with Tyr42 in contact with water molecules. The barrier is shallower in the direction of the N$_{water}$(HC4)-axis, indicating a mechanism for the solvent exposure of the chromophore: Partial hydration of the chromophore binding pocket interferes with the strong hydrogen bond between the chromophore and Tyr42. Once it is broken, full solvent exposure of the chromophore occurs.

Glu46 is located in a helix comprising residues 42 to 50. This helix forms a bundle with the helical conformation of residues 53-59. On average, 4 to 5 hydrogen bonds exist between the protein backbone atoms of residues 42 to 59, in their folded conformation. Upon the solvation of Glu46, these helices undergo a conformational change. Fig. 4.5(c) displays the free energy profile as a function of the number of backbone hydrogen bonds of residues 42 to 59 and the number of water molecules within a radius of 0.35 nm around Glu46. The profile has a minimum at 4.5 backbone hydrogen bonds and zero water molecules around Glu46, denoted as I. Going into the direction indicated by label A, an increase of water molecules around Glu46 occurs, while the number of backbone hydrogen bonds drops. Although this is not a minimum, it clearly indicates the loss of helical conformation upon the solvent exposure of Glu46.

These results suggest that conformational rearrangements occur in the receptor state of PYP, with the solvent exposure of the chromophore being the predominant process. Low-temperature spectroscopy of PYP indicates that heterogeneity in the ground state population causes broadening of the PYP absorption spectrum, leading to two different photochemical pathways upon excitation [114–116]. The existence of a PYP conformation with a solvent-exposed chromophore would fit with these observations. Since the chromophore in water has a pKa close to 9, chromophore protonation would be feasible in such a solvent exposed conformation. However, pH titration of the formation of pB-dark shows no change in the shape of the spectrum until low pH, indicating a pKa of 2.7 for protonation of the chromophore [117]. Nevertheless, a closer examination of this titration curve using an advanced fitting procedure hints at the existence of a small fraction of PYP with a solvent exposed chromophore (J. Hendriks, personal communication). Finally, the hydrogen bond between the chromophore and Tyr42 is very short and may contain covalent characteristics [118]. In the simulation, electrostatic and van der Waals interactions describe hydrogen bonds, possibly underestimating the strength of this particular hydrogen bond. A stronger hydrogen bond between Tyr42 and the chromophore would reduce the degree of solvent exposure of the chromophore. The observation that water molecules have access to the chromophore in the receptor state would suggest the existence of other proton donors besides Glu46 for the chromophore protonation reaction later in the photo cycle.

Recovery of the receptor state

The recovery of the receptor state pG from the signaling state pB requires multiple conformational and chemical rearrangements. Before discussing our investigation of the recovery reaction, we first focus on the pB′ simulation that sampled the formation of the signaling state, figure 4.6(a). This figure shows the free energy profile as a function of the distance between the chromophore and the side chain of Tyr42, and the distance between the side chains of Tyr42 and Glu46. The pB′ system, figure 4.6(a), starts with an intact chromophore binding pocket, indicated by the well at $d_{Chro-Tyr42}$ = 0.6 nm, $d_{Tyr42-Glu46}$ = 0.5 nm, labeled I. During the pB′ simulation, the chromophore and Glu46 become solvent exposed, as indicated by the series of free energy wells towards larger distances between the chromophore and Tyr42 (label II), and larger values for the Tyr42-Glu46 distance (label III).

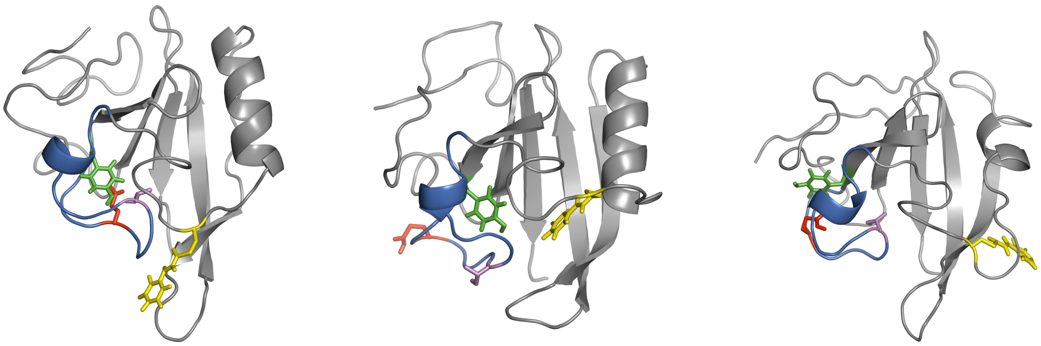


(a) pB′

(b) pB^{deprot}

(c) pG′

Figure 4.6: **Free energy profiles of the chromophore binding pocket.** The plots are drawn as a function of the distance between the centers of mass of the phenolate ring of the chromophore (HC4) and the side chain of Tyr42 (horizontal axis) and the distance between the centers of mass between the side chains of Tyr42 and Glu46 (vertical axis) for the parallel tempering simulations representing (a) pB′, (b) pB^{deprot} and (c) pG′. The star in (a) indicates the starting configuration of the pB^{deprot} and pG′ systems. In (c) the free energy profile of pB′ (a) is drawn in light blue. The labels refer to free energy minima.



(a) IV

(b) II

(c) III

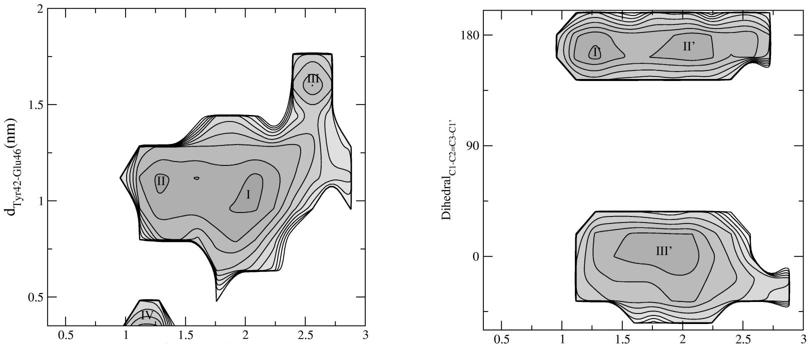
Figure 4.7: **Snapshots from the pG′ simulation.** The ribbon representation shows the protein backbone in grey, with residues 42-59 highlighted in blue. The stick models display the side chains of residues Tyr42 (green), Glu46 (red), Thr50 (pink) and Cys69-HC4 (yellow).

This graph shows that the formation of the signaling state is not complete, since eventually both groups become solvent exposed, as we have shown before [104]. The star indicates a signaling state conformation obtained from our previously published pB simulation that we used to initiate two parallel tempering simulations, pB^{deprot} and pG′. In this conformation, both the chromophore and Glu46 are solvent exposed. In the pB^{deprot} and pG′ simulations, the chromophore and Glu46 are in receptor-like protonation states: the chromophore is deprotonated and Glu46 is protonated. The systems differ in configuration of the chromophore: The pB^{deprot} system contains a *cis*-oriented chromophore. pG′ contains a *trans* chromophore and differs only in conformation from the receptor state. Table 4.1 lists the details for the topologies of the systems.

Figures 4.6(b) and 4.6(c) show the free energy profiles of the chromophore binding pocket, as sampled in the pB^{deprot} and pG′ simulations, respectively. For pB^{deprot} the free energy profile is relatively straightforward: it contains a minimum at d$_{HC4-Tyr42}$ = 1.9 nm, d$_{Tyr42-Glu46}$ = 1.0 nm, labeled I. More features emerge in the free energy profile sampled during the pG′ simulation. The well indicated by I resembles the minimum in the pB^{deprot} simulation. In addition, the free energy profile contains minima at (d$_{HC4-Tyr42}$ = 1.3 nm, d$_{Tyr42-Glu46}$ = 1.3 nm), labeled II, at (d$_{HC4-Tyr42}$ = 2.6 nm, d$_{Tyr42-Glu46}$ = 0.6 nm), labeled III. These wells respectively represent conformations with the chromophore oriented toward the protein interior (II) and conformations with a hydrogen bond between Tyr42 and Glu46 (III). Finally, the free energy profile contains a feature (label IV) at hydrogen bonding distance for Glu46 and Tyr42, and at a distance around 1.5 nm between Tyr42 and the chromophore. This feature represents conformations with a partially reformed chromophore binding pocket. Interestingly, the wells II and III in the free energy profile of pB′ overlaps with the free energy profile of pG′ (the light blue lines in figure 4.6(c), indicating that the pG′ simulation is actually sampling part of the recovery reaction.

Figure 4.7 shows three snapshots from the pG′ simulation representing the free energy minima indicated with a roman numeral in figure 4.6(c). Snapshot IV shows a partially refolded chromophore binding pocket, with a hydrogen bond between Tyr42 and Thr50. Oriented toward the protein interior, Glu46 is alongside Tyr42. Also oriented toward the protein interior, the chromophore is close to these residues that make up the chromophore binding pocket, but does not interact with them. Residues 50-55 form a helix in the receptor state structure, but are now in a coil conformation and prevent the chromophore from entering its binding pocket. Snapshot II shows that the chromophore can enter the protein interior and approach Tyr42. Nevertheless, Glu46 is far from the protein interior, interacting with residues in the N-terminal cap. Moreover, Thr50 has shifted toward the protein surface and interacts with solvent molecules. The pG′ simulation also sampled conformations with Glu46 inside the protein, as snapshot III displays. Such conformations are far from the folded receptor state, since both the chromophore and Tyr42 are solvent exposed and oriented to opposite sides of the protein. In these conformations, the region containing Glu46 blocks the reorientation of the chromophore towards the protein interior.

Comparison of the pB^{deprot} and the pG′ simulations would suggest that the *cis*/*trans* configuration of the chromophore significantly affects the conformational space accessible to the protein. To test this dependence we performed a parallel tempering simulation, indicated as pB$^{deprot}_{flexible}$, with a lowered rotation barrier for the chromophore double bond. Figure 4.8(a) shows



(a) Chromophore binding pocket (b) Cis/trans

Figure 4.8: **Free energy profiles of the pB$^{deprot}_{flexible}$ simulation.** The plots are drawn as a function of the distance between the centers of mass of the phenolate ring of the chromophore and the side chain of Tyr42 (horizontal axis) and the distance between the centers of mass of the side chains of Tyr42 and Glu46 (vertical axis, (a)) or the dihedral angle along the chromophore double bond (vertical axis, (b)). The dihedral is measured along the double bond. The labels indicate free energy minima. Note that the free energy wells in (a) and (b) are not related.

the free energy profile as a function of the distance between the chromophore and Tyr42, and Glu46 and Tyr42. This profile has two minima separated by 1 k_BT at d$_{Tyr42-Glu46}$ = 1.1 nm, with chromophore-Tyr42 distances of 1.4 nm (I) and 2.1 nm (II). This broad basin is also present in the pB^{deprot} and pG′ simulations. The pB$^{deprot}_{flexible}$ profile contains two additional minima, of which one is at a barrier of 4 k_BT from the broad basin, with (d$_{Chro-Tyr42}$ = 2.6 nm, d$_{Tyr42-Glu46}$ = 1.7 nm), labeled III. Minimum IV is at (d$_{Chro-Tyr42}$ = 1.2 nm, d$_{Tyr42-Glu46}$ = 0.4 nm), representing a conformation closely resembling snapshot IV from the pG′ simulation. These two minima are extra features in comparison to the pB^{deprot} simulation. Plotting the free energy profile as a function of the chromophore-Tyr42 distance and the dihedral along the chromophore double bond, figure 4.8(b) shows *cis* conformations at +10^o and -20^o (III′), and *trans* conformations at 172^o (I′ and II′). The extra well (label I′) appears only for the *trans* configurations, indicating that the isomerisation of the chromophore affects the conformational space accessible to the protein. Minimum I′ at hydrogen bonding distance for the chromophore and Tyr42 indicates that the *trans* configuration is required for the reformation of the receptor state. This observation suggests that the recovery of the receptor state, after triggering the *cis* to *trans* isomerisation of the chromophore, would consist of the refolding into a helical conformation of the region around Glu46. These results are consistent with the observation that the recovery reaction is accelerated via light-induced isomerisation of the chromophore [110].

Concluding remarks

Previous simulations have shown that formation of the signaling state of PYP is characterized by the solvent exposure of the chromophore and Glu46 [104]. These results are in agreement with NMR spectroscopy data obtained from Δ_{25}-PYP [102]. However, the two conformations do not compare well in all aspects: The NMR structure has a more extended conformation for the central β-sheet and a larger degree of unfolding in the α-helical region containing Glu46. In a simulation of the truncated mutant protein we observed that solvent exposure of the chromophore and Glu46 occur more easily, and that such unfolded conformations compare well with the NMR structure. This validates our previous prediction of the signaling state, as the presence of the N-terminal cap restricts the conformational space of the sensing core of the protein. Furthermore, this validation provides insight into the role of the N-terminal domain. By restricting the conformational freedom of the central β-sheet, the N-terminal cap limits the pathways for and the extent of light-triggered unfolding.

To gain more insight into the conformational characteristics of the receptor state, we performed a parallel tempering simulation. Previously we have shown that in the receptor state, the chromophore binding pocket assumes different conformations [104]. Here, we have demonstrated that the chromophore is in contact with the solvent, even in the receptor state at a small fraction of time. Also Glu46 can become solvent exposed, leading to loss of α-helical conformation. Solvent exposure of Glu46 is less likely to occur than solvent exposure of the chromophore. The simulation of pG sampled several chromophore solvent exposure events. Such conformations with a solvent exposed chromophore would explain the heterogeneity observed in the low-temperature absorbance spectra. Moreover, these conformations would indicate that solvent molecules, to a lesser extent than Glu46, can act as proton donors for the protonation of the chromophore during the PYP photo cycle.

Studying the recovery process of PYP with molecular simulation techniques is a huge challenge, due to the time scales involved. In an attempt to sample the recovery reaction of the receptor state, starting from a partially unfolded representative conformation of the signaling state, we present parallel tempering simulations initiated from different configurations along the recovery reaction pathway. Complete refolding did not occur in any of the recovery simulations, but we observed overlap in the free energy profiles of a simulation starting from an unfolded conformation and a simulation of the folded conformation. In the simulation with a deprotonated *trans*-chromophore and protonated Glu46, the chromophore did enter the protein interior. Also, the hydrogen bond between Tyr42 and Glu46 reformed, but these events did not occur simultaneously. The simulations show that the *cis* or *trans* configuration of the chromophore has a significant effect on the ability of the protein to refold into its receptor state. A simulation with a lowered rotational barrier of the chromophore double bond further substantiated this conclusion.

Chapter 5

Protein Triads: A new method to analyse collective motions in proteins*

Ensembles of protein structures, generated in various ways, assemble in large and complex datasets. It is not straightforward to find the motions that underly the functional mechanism of a protein from such a dataset, since it contains all motions available to the protein (within the boundaries of the sampled conformational space). These motions comprise many modes, such as overall translation and rotation, breathing, thermal flexibility, side chain rotation and segmental flexibility. In the dynamics of signal transduction, it is the general view that sensor proteins undergo conformational changes that affect large parts of the molecule.

Discovering the general trends in such data requires data extraction techniques. When viewing motions in molecular systems, three components express variation: time evolution, particles involved in the motion and direction of the motion. Currently, the principal component analysis method known as Essential Dynamics is generally applied to extract the large concerted fluctuations from conformational datasets. Such models explain the variance in time and a mixed mode containing the particles and direction. In this work we present a new method to extract internal protein motions: Protein Triads. This method aims to explain the variance in the three components required for understanding collective protein motions.

Using a simple atomic model we show that Protein Triads performs equally well in capturing the variance in the data sets as Essential Dynamics. Applying both methods to a parallel tempering simulation of Photoactive Yellow Protein (PYP) shows that Protein Triads is able to separate relevant motions into different model components, resulting in a better interpretable model. A direct consequence is that it is easier to visualize the extracted motion.

PYP is a member of the PAS protein family, that is involved in signaling and protein-protein interactions. Using molecular dynamics simulations of the PAS folds in HERG, the N-terminal domain of a human potassium channel, and FixL, a bacterial oxygen sensor, we employ Protein Triads to compare the dynamics exhibited in the three protein systems in a qualitative and a quantitative way. The latter enables comparison of dynamics in different protein systems.

---

*This chapter is based on J. Vreede, K.J. Hellingwerf, A.K. Smilde, W. Crielaard and H.C.J. Hoefsloot *Protein Triads: A new method to analyse collective motions in proteins* Submitted to J. Comput. Biol.

Introduction

The general consensus in structural and molecular biology is that the function of a protein is outlined in its structure. The efforts to increase our insight into protein structure at the atomic level are reflected in the still growing number of new deposits in the Protein Data Bank [119]. Unfortunately, our understanding of protein function and protein-protein interactions has not increased concomitantly. Recently, it has become more evident that internal protein dynamics play a fundamental role in protein function [120]. In the case of signal transduction proteins, this insight is already biochemically substantiated [121]. The detection and communication of a signal require (transient) conformational changes and these are confined within the conformational freedom of a protein. Via these structural restrictions, its dynamic characteristics are defined within itself. Now the challenge is to be able to characterize, predict, compare and understand these internal dynamics from a protein structure as such. Models describing the functional activity of a protein should consider both structural and dynamic details.

Proteins in general can be classified into families on the basis of structural similarity. Within these protein families, all members share the same typical fold. One of these structural families, the PAS-fold family, has been very useful in the strive to understand functional protein dynamics [20]. PAS is the abbreviation of PER-ARNT-SIM, the first three proteins revealing the PAS sequence [23]. This family brings together folds that were previously recognized as LOV and/or GAF domains and now contains more than 1100 members [21]. PAS domains play a key-role in signal transduction toward the regulation of a variety of processes in representative organisms from all domains of life, including *Homo sapiens* [10, 15, 122]. Several detailed studies on one of the members of the PAS domain family, the Photo-active Yellow Protein (PYP) [20], have indicated that when this protein passes through its signal transduction cycle distinct conformational changes occur [25, 26, 29, 30, 105, 123]. In previous work we have explored the idea that similar structures induce a similar dynamic behaviour in a protein familiy.

Molecular simulation techniques sample protein conformational space, resulting in large amounts of data that are difficult to interpret in terms of functional motion. Up to now, a *two-way* method known as Essential Dynamics (ED) is predominantly used to extract large correlated fluctuations from sets of structures (*i.e.* a trajectory) [45]. ED is the application of principal component analysis (PCA) to protein dynamics and it derives a model for structural variation. Using a combination of CONCOORD simulations and ED/PCA, common characteristics in the flexibility of different PAS folds could indeed be identified [109]. Only the use of a subset of atoms, identical in size for each protein domain, allowed direct comparison between the various PAS proteins. Unfortunately, the construction of a subset representing the family fold biases the analysis to finding similar flexible modes. Another study [124] performed a qualitative comparison of the conformational flexibilities exhibited by PAS domains. Using Molecular Dynamics (MD) simulations and ED/PCA the authors concluded that PAS domains have a limited number of general solutions for sensing and signaling. From these studies it became clear that a new methodology needs to be explored in order to describe and compare functional dynamics in proteins.

In the current study we will evaluate Tucker-3 models as a method to understand functional protein dynamics [125]. This type of model has been developed by Ledyard R. Tucker [126, 127] and in contrast to ED/PCA, arranges dynamics-data as a three-dimensional matrix and subsequently uses three-way analysis techniques to structure the results. Three-way analyses, including Tucker-3, have, to our knowledge, never been applied to protein dynamics before. We call the application of Tucker-3 to protein dynamics *Protein Triads*.

Applying both ED/PCA and Protein Triads to a simple system containing eleven atoms, and a parallel tempering simulation of PYP, we will show that comparable motions are identified by both methods. In addition, we use Protein Triads to compare internal protein dynamics of three members of the PAS protein family, facilitated by MD simulations of two PAS domains, HERG, the N-terminal part of a human potassium channel, and FixL, a bacterial oxygen sensor, and the parallel tempering simulation of PYP [104]. A qualitative comparison shows that the three PAS domains exhibit different flexible modes. Using the directional modes, Protein Triads facilitates direct quantitative comparison.

Theory

A dataset describing the trajectories of all individual atoms within a protein is often large and complex. The construction of a model is a way to increase the comprehensibility of the data. A good model must be less complex than the original data, but must still contain the most important aspects of the data. To understand how data analysis techniques are able to extract relevant variance from a set of protein structures, we need to fully understand what a motion is comprised of. Motion is considered as the change in position of an object in time. However, the moving bodies making up the internal protein motions are linked to one another via inter-atomic bonds and cannot be considered separately. Consequently, understanding internal protein motion requires three aspects: which atoms are moving, the time profile, and the direction of the motion.

The principal component analysis (PCA) in Essential Dynamics is commonly explained in terms of a covariance matrix of the trajectories. The results, *i.e.* the essential eigenvectors, are a model for the data. However, to allow a comparison between PCA and Tucker-3, the two methods must be explained in terms of the minimization of the differences between the model and the data. For PCA, the model parameters result from the solution of the following problem:

$$\underset{\boldsymbol{T},\boldsymbol{P}}{min} \parallel \boldsymbol{X} - \boldsymbol{T}\boldsymbol{P'} \parallel^2 \tag{5.1}$$

Here $\boldsymbol{X}$ is the matrix of size $NK \times J$ that contains the data. The number of rows is equal to the number of atoms (N) times the number of spatial directions (K). The number of columns is equal to the number of sample-points (J). Symbol $\boldsymbol{T}$ holds the scores. These are comparable to the eigenvectors in a covariance matrix. The size of $\boldsymbol{T}$ is $NK \times R$. The columns of $\boldsymbol{P}$ are called the loadings. The size of $\boldsymbol{P}$ corresponds to the number of time-points J times the number of principal components R. In ED these are explained as the projection of the original trajectory onto the eigenvectors. $\boldsymbol{P'}$ is the transpose of $\boldsymbol{P}$. ($\parallel \boldsymbol{A} \parallel^2$) is the squared Euclidean norm of $\boldsymbol{A}$. The PCA model of the data is graphically displayed in figure 5.1 (top). A larger number of principal components yields a better model prediction but results in a larger and more complex model. Consequently, the choice of principal components is a trade-off between complexity and predictive power of the model. In ED the principal components are called the essential

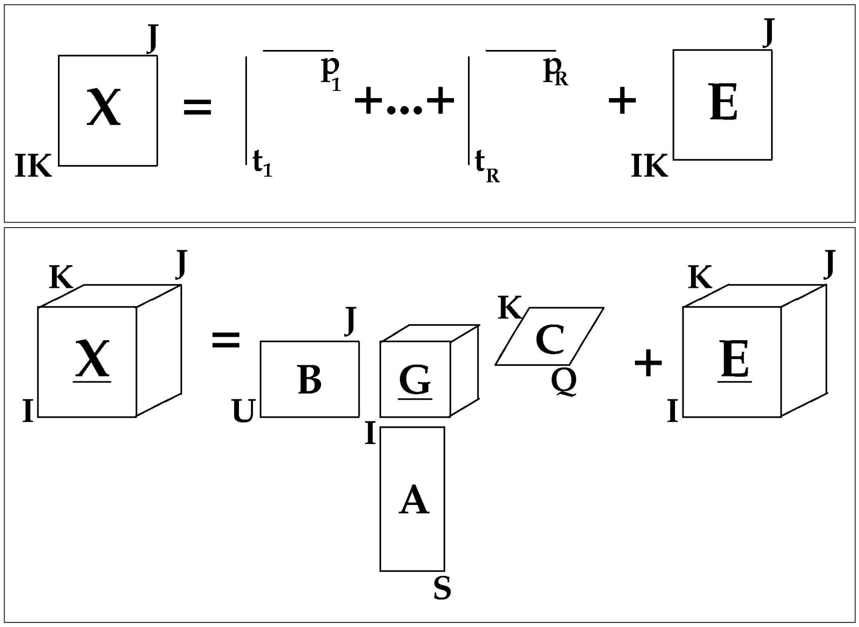


Figure 5.1: **Schematic drawing of the models for protein dynamics as generated by (top) ED/PCA and (bottom) Tucker-3.** The two methods are based on the minimization of the error matrix $\boldsymbol{E}$ or array $\underline{\boldsymbol{E}}$. The model derived by ED/PCA is composed of R components that consider the temporal aspects (t) of protein dynamics along with atomic fluctuations (p). The Tucker-3 model decomposes protein motions in temporal ($\boldsymbol{B}$), atomic ($\boldsymbol{A}$) and spatial components ($\boldsymbol{C}$) that are linked through the core array $\underline{\boldsymbol{G}}$. The values for S. U and Q represent the number of components $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ respectively. I, J and K indicate the size of the original data matrix/array.

eigenvectors. The prediction of a single data point in the model is defined as follows:

$$\hat{x}_{ij} = \sum_{r=1}^{R} t_{ir} p_{jr} \tag{5.2}$$

The element t_{ir} is on row number i and column number r of $\boldsymbol{T}$ and p_{jr} is element j,r of $\boldsymbol{P}$. The physical interpretation of the columns in $\boldsymbol{P}$ in the context as presented here is a time series. The scores $\boldsymbol{T}$ represent vectors in the NK configuration space. To simplify the interpretation of directions in the NK dimensional space, we advocate the Tucker-3 method to analyze trajectory data. The resulting model is less complex and the interpretation of the model parameters is straightforward. Solving problem 5.3 gives the model parameters for a Tucker-3 model:

$$\underset{\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{G}}{min} \parallel \boldsymbol{X} - (\boldsymbol{A} \otimes \boldsymbol{C})\boldsymbol{G}\boldsymbol{B}' \parallel^2 \tag{5.3}$$

Here $\boldsymbol{X}$ is a rearrangement of the three-way array $\underline{\boldsymbol{X}}$ of a size defined by the number of atoms N × number of spatial directions K × number of time-points J. After rearrangement of $\underline{\boldsymbol{X}}$ the dimensions of $\boldsymbol{X}$ are $NK \times J$. The columns of $\boldsymbol{A}$ give the relative amplitudes of the motions of the atoms under consideration. The columns in $\boldsymbol{C}$ represent vectors in Cartesian space and indicate a direction. $\underline{\boldsymbol{G}}$ is the so-called core array that connects the three matrices. Similar to $\underline{\boldsymbol{X}}$ it is rearranged to $\boldsymbol{G}$ with dimensions $SQ \times U$. The columns of matrix $\boldsymbol{B}$ are time-series, similar to $\boldsymbol{P}$ in ED/PCA. $(\boldsymbol{A} \otimes \boldsymbol{C})\boldsymbol{G}$ represents $\boldsymbol{T}$ in ED/PCA. This shows that the Tucker-3 model is a restricted version of PCA. The restrictions improve the interpretability of the model. A schematic drawing of the matrix multiplication is given in figure 5.1 (bottom). The symbols S, U and Q determine the size of the matrices that construct the model. Larger values for S, U and Q give better approximations. The Tucker-3 model can also be considered as a summation of triads, as depicted in figure 5.2. Each data point is predicted by:

$$\hat{x}_{ijk} = \sum_{s=1}^{S}\sum_{u=1}^{U}\sum_{q=1}^{Q} a_{is} b_{ju} c_{kq} g_{suq} \tag{5.4}$$

A triad is comprised of three vectors (a_s, b_u, c_q) and a scalar g, where a_s is the s-th element of $\boldsymbol{A}$, b_u the u-th element of $\boldsymbol{B}$ and c_q the q-th element of $\boldsymbol{C}$. The scalar is an element of the core array $\underline{\boldsymbol{G}}$. Each element in $\underline{\boldsymbol{G}}$ is a measure of the relative weight of the triad under consideration. The explained variance is relative to the square root of the scalar g, if $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ are orthogonal. The spatial vectors in the matrix $\boldsymbol{C}$ give the direction of the motion in a triad. Similarly, the temporal vectors in $\boldsymbol{B}$ describe time-series. The vectors in the atom mode $\boldsymbol{A}$ indicate the relative amplitudes of the individual particles. Consequently, each component in the Tucker-3 model directly relates to different aspects of an internal protein motion, as it is composed of the displacement of an atom as a function of time, relative to its neighboring atoms. Although Tucker-3 requires more matrices than ED/PCA, due to the decomposition of $\boldsymbol{T}$, it is clear that all parameters in a Tucker-3 model have a simple and physical interpretation.

Understanding the dynamics underlying protein function requires the extraction of collective motions from the large and complex datasets resulting from sampling protein conformational diversity. Improving the interpretability of models that aim to explain the variance in these datasets also allows better insight in protein dynamics. Therefore, the Tucker-3 model, in comparison to ED/PCA, improves the interpretation of the large concerted motions exhibited by proteins. We propose in analogy to Essential Dynamics (using PCA to analyze protein dynamics) to call the application of Tucker-3 to analyze protein dynamics *Protein Triads*.

Computational details

To facilitate a comparison between ED/PCA and Protein Triads, we used four systems. First, the comparison focuses on a simple system that contains eleven atoms undergoing a hinge bending motion. Second, increasing complexity, we used a previously published parallel tempering simulation of PYP to compare ED/PCA and Protein Triads [104]. Finally, we performed Molecular Dynamics (MD) simulations of two members of the PAS family, HERG, the N-terminal domain of a human potassium channel, and FixL, a bacterial oxygen sensor. Together with the dataset on PYP, these simulations served as input for Protein Triads for a comparison of their internal dynamics.

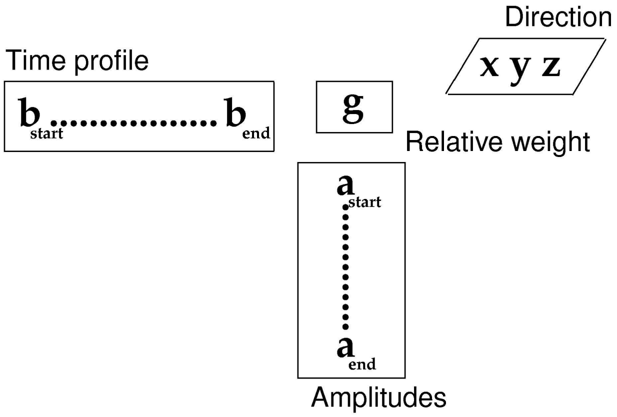


Figure 5.2: **Schematic drawing of the Tucker-3 model as composed of triads.** A triad is the combination of an atomic mode with a temporal mode and a spatial mode. a and b represent an atom mode and a time mode respectively. x, y and z represent a spatial mode in Cartesian coordinates. The value for g gives the relative weight of the triad.

Setup of the MD simulations

Starting structures for the MD simulations originated from the Protein DataBase, using 1BYW for HERG and 1DRM for FixL. Both structures missed side chain atoms at the protein surface and required a simple fitting procedure to complete the structures prior to the simulation setup. The C-terminal helix in the FixL structure, comprising residues 257-270 was removed, as it is not part of the PAS fold. Each starting point was placed in a periodic box, followed by the addition of polar and aromatic hydrogen atoms to accommodate standard protonation states. SPC water molecules [90] filled the box at a density of 1.0 kg/l, those within a distance of 2.6 $\AA$ of the protein atoms and those inside the protein were removed. Crystal waters remained, if present. To neutralize a positively charged system, chloride ions replaced water molecules at the most electropositive positions. Similarly, replacement of water molecules at the most electronegative positions by sodium ions neutralized a negatively charged system. Internal strain in the system was relaxed by a number of steps of the conjugated gradient method, followed by a two-step equilibration procedure. First, the positions of the water molecules and hydrogen atoms were relaxed by the application of harmonic position restraints to the heavy atoms of the protein. After 10 picoseconds, the constraints were released to equilibrate the whole system for 100 ps. Equilibration took place at a constant pressure of 10^5 Pa and a constant temperature of 300 K. The systems were then sampled at a rate of 1 ps in the isothermal-isobaric ensemble running multiple replicas using different starting velocities. The application of the LINCS [95] algorithm for covalent bonds in the protein and the SETTLE [96] algorithm for the water interactions allowed for a time step of 2 femtoseconds. To maintain the NpT ensemble, the Parrinello-Rahman barostat [128, 129] and the Nose-Hoover thermostat [130, 131] were used. The simulation setup and sampling were performed with the GROMACS software package [91, 132]. Interactions between atoms were described by the GROMOS96 force field [133, 134]. Aliphatic hydrogens were treated implicitly using the united atom approach, while polar and aromatic hydrogens were defined explicitly. Electrostatic interactions were resolved using the Particle-Mesh-Ewald summation method, while van der Waals interactions were treated with a cut-off at 9 Å. For HERG, three MD runs were performed and for FixL four runs were initiated, all runs starting with differently generated velocity distributions corresponding to a temperature of 300 K. In total, the three HERG-replicas covered 60 ns and the four FixL replicas covered 50 ns.

Analysis

The combined molecular dynamics trajectories of FixL and HERG, at a rate of one snapshot per 10 ps, served as input for the analysis procedure. The parallel tempering trajectory contained snapshots of every ps. Prior to data analysis, each trajectory frame was aligned with the equilibrated starting structure to remove overall rotation and translation using a least squares fitting procedure.

After structural alignment the coordinates for each atom were centered around its average value. The ED/PCA analysis was performed in GROMACS [132]. MATLAB was used for the Tucker-3 analysis in combination with the N-way toolbox from Anderson and Bro [135]. Both analyses included only the data concerning the C_α atoms, since our interest is focused on large backbone fluctuations rather than side chain movements. The 'simple system' was created in MATLAB. Protein graphics were created with Molscript [136] and Raster3d [137].

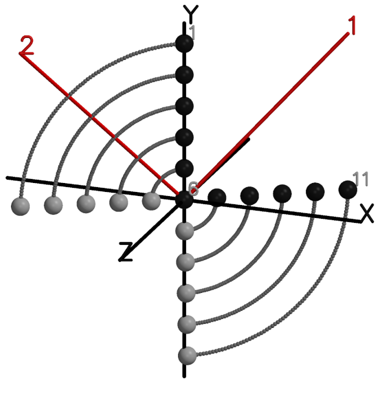


Figure 5.3: **Graphical representation of a simple system containing 11 atoms.** The spheres indicate the initial (black) and final (grey) positions of the atoms in an xyz coordinate system. Motion occurs along the lines connecting the black and grey spheres. The red axes indicate the first and second spatial modes.

Results and discussion

In the analysis of protein dynamics using data analysis techniques the composition of the data must be understood clearly. The data from a molecular dynamics simulation is arranged in snapshots that contain the Cartesian coordinates of each atom at subsequent time points. ED/PCA considers such a data-arrangement as a two-dimensional matrix. The two dimensions represent time and the atomic positions in Cartesian space: the $3N$-dimensional configuration space. As stated in the Theory section, an atomic motion inside a protein is composed of three aspects. Using three dimensions is a more natural way to represent the data, separating the atom index from the positional variation. The analysis of a three-dimensional matrix requires three-way analysis techniques, such as Tucker-3. We propose in analogy to Essential Dynamics (using PCA to analyze protein dynamics) to call the application of Tucker-3 to analyze protein dynamics *Protein Triads*.

Comparing ED/PCA and Protein Triads: a simple system

Before starting on the complex dynamics of protein systems, we first want to elucidate the differences between ED and Protein Triads by considering a simple system containing eleven atoms. Figure 5.3 shows the initial and final positions of the atoms in an xyz coordinate system. Initially, the first five atoms are positioned on the y-axis. Atom 6 is located at the origin. $2z = 3x$ describes the location of atoms 7 to 11. The line connecting atoms 1-6 rotates 90° and the line through atoms 6-11 rotates -90°, so atom 6 does not change position. The final positions of atoms 1 to 5 are described by $2z = 3x$, and the last five atoms end on the y-axis. The red axes in figure 5.3 indicate the direction of the movement. From start to end, 80 snapshots of intermediate configurations make up the raw data for ED/PCA and Protein Triads analysis.

ED/PCA captures 100 % of the motion in two eigenvectors. Protein Triads using two modes for atom, time and spatial directions, also captures 100 % of the motion. The methods do not differ in amount of explained variance. Using the amount of explained variance per component as a sorting criterium, the first component is the most important one. The core array in Protein Triads determines the order of the (orthogonal) components, since the explained variance of a component is equal to its corresponding value in the core array, squared.

g_{111} = -56.543 g_{112} = -0.115 g_{121} = 0 g_{122} = 0
g_{211} = 0 g_{212} = 0 g_{221} = 0.555 g_{222} = -11.667

Table 5.1: **Core array G of the Protein Triads analysis of the simple system**

Figure 5.4 displays the results of analysing the motions in the system, using ED/PCA and Protein Triads. The time modes in both methods, figure 5.4(a) and (c), exhibit similar features, apart from a difference in sign in the second mode. Besides temporal information, the models contain atom-related modes and it is in this interpretation that the models differ. In ED/PCA the vectors describe the individual differences in position between the atoms, figure 5.4(b) solid lines. Interpreting this figure is not straightforward. Averaging over the three directions gives an indication of positional difference per atom, figure 5.4(b) dashed lines, and improves the

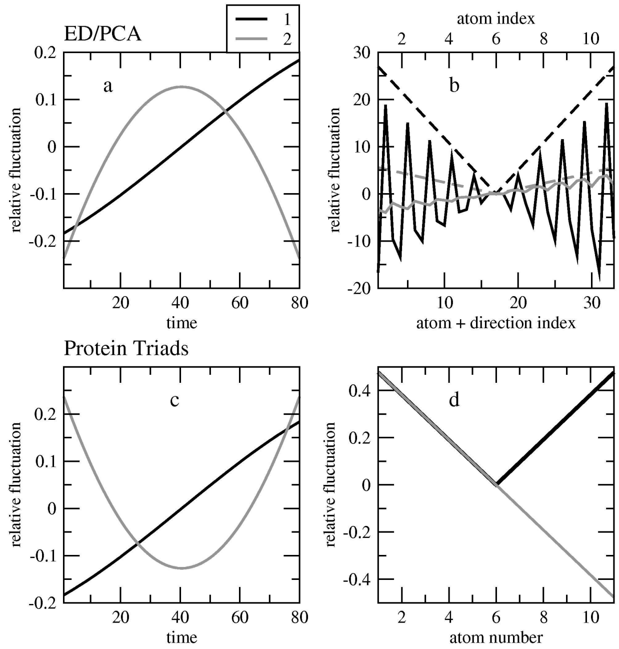


Figure 5.4: **ED/PCA and Protein Triads analysis of the simple system** The upper graphs display the results from the ED/PCA analysis: (a) the time components and (b) the atom-directional components, with the dashed lines representing the root mean square of the atom-directional components. The lower graphs show the results from the Protein Triads analysis: (c) the time components and (d) the atom components.

interpretability of the vectors. Both components show a linear decrease in positional difference for atoms 1-6, with a value of zero for atom 6. Atoms 7 to 11 exhibit increasing fluctuations.

Protein Triads separates direction from the atom index. These directional modes are orthogonal vectors in three-dimensional space and highlighted as red lines in figure 5.3. The first spatial mode describes the movement of the atoms, and the second indicates the direction in which the atoms move. Figure 5.4 (d) displays the atomic modes in the Protein Triads model. In both atomic modes the middle atom, 6, has zero amplitude, indicating that it doesn't move during the collective motion. The first atomic mode shows a linear decrease of motion for atoms 1 to 6, increasing again in the opposite direction for atoms 7-11. This reflects the linearly decreasing respectively increasing velocities of these atoms. Also, the component clearly shows that the motions of atoms 1 and 11, 2 ad 10 etc. are correlated, since these pairs have similar values. The second atomic mode shows a decrease in the amplitude for atoms 1-5 and atoms 11-7, in that order, indicating anti-correlated behaviour for the atom pairs. Both methods indicate a correlation between atoms 1 and 11 up to atoms 5 and 7. The improvement that Protein Triads offers over ED/PCA is that the atomic modes clearly indicate the atoms involved in the correlated and anticorrelated motions.

Comparing ED/PCA and Protein Triads: Photoactive Yellow Protein

Formation of the signaling state of PYP is initiated by photo-excitation of the chromophore, followed by proton transfer in the chromophore binding pocket. This process results in a negative charge on glutamic acid 46, in the protein interior. With this situation as a starting point, a previously published parallel tempering simulation of PYP [104] sampled the formation of the signaling state, where Glu46, in search for stabilization, is the driving force of the partial unfolding of the protein. Eventually both Glu46 and the chromophore become exposed to solvent outside the protein core. The trajectory contains snapshots of the sampled conformational space of the protein. Each sample point comprises the three dimensional positions, defined in Cartesian coordinates, of the C_α-atoms. As such, it is the input for ED/PCA and Protein Triads analysis, facilitating a comparison between the two methods. First, we analysed the variation in the PYP data using the essential dynamics technique (see table 5.2 for a summary of the results). In general, the first few of the eigenvectors spanning the conformational space are defined as "essential" [46]. If an ED/PCA model is used to fully cover the data variance, the number of components would be equal to the dimension of the conformational space. The number of eigenvectors to choose as essential depends on the change in eigenvalue, since the eigenvalue relates directly to the explained variance of that component. A model of 5 essential eigenvectors explains 50.6 % of the variance in the data. After sorting on eigenvalue, the first two eigenvectors capture one quarter of the conformational flexibility.

ED/PCA

Component	Expl.var.(%)	Eigenvalue (nm^2)
1	16.7 (16.7)	1.69
2	10.6 (27.3)	1.07
3	9.8 (37.1)	1.00
4	8.0 (45.1)	0.81
5	5.5 (50.6)	0.56

Protein Triads

Component	Atom index	Time	Direction	Expl.var.(%)	Value core array g
1	1	1	1	14.8 (14.8)	-109
2	1	2	2	5.6 (20.4)	67.2
3	2	3	3	5.3 (25.7)	-65.5
4	3	4	1	5.3 (31.3)	-65.3
5	4	5	2	2.5 (33.8)	44.4

Table 5.2: **Details of the ED/PCA analysis and the (10,10,3) Protein Triads analysis on PYP.** The first five components are listed, sorted by explained variance. The explained variance (expl.var.) is given in percentages, with the summed variance in parentheses. For each eigenvector (ED/PCA) the eigenvalue is given. For each triad (Protein Triads) the composition of modes from the atom index, the time and the direction is listed. The atom index indicates which atoms are moving in a correlated way. The values from the core array are given for each triad also. The core array assigns a weight factor to each combination of the atom index, time and direction.

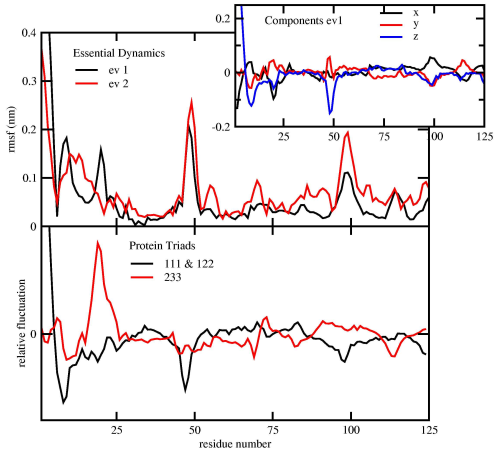
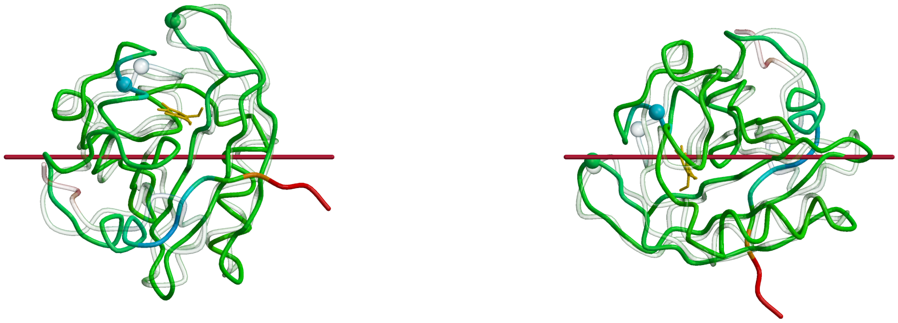


Figure 5.5: **Atomic variation in PYP in the first few model components.** (top) Atomic fluctuations in the first and second eigenvector from ED/PCA. (bottom) Relative amplitudes of the first three triads. The inset shows the fluctuations in the x, y and z direction of the first eigenvector. The atomic variation in the eigenvectors is displayed as the root mean square fluctuation in nm. The relative atomic amplitudes in Protein Triads have no unit.



(a) spatial mode 1

(b) spatial mode 2

Figure 5.6: **Graphical representation of the PYP triads** The red axis indicates the first (a) and second (b) spatial mode as identified by Protein Triads analysis. Orientation of the drawings is such that the spatial mode is parallel with the viewing plane. The ribbon drawings represent PYP in typical conformations that optimally display the amplitude of the motion, with the transparent and the colored coils respectively representing the highest and lowest values in the relative fluctuation in the respective time profiles. The colors indicate the relative atomic fluctuation from blue (lowest value) to red (highest value) via green and yellow. Effectively, this means that the red and the blue parts move in an anti-correlated way. The spheres highlight residues 47 (blue) and 98.

A first prerequisite when using Protein Triads is establishing the correct number of components. Avoiding loss of explained variance, it is the aim to have a minimal number of modes. Since each atom has three directions available, the spatial mode requires no more and no less than three modes. For the temporal mode and the atom index, the number of modes may be varied. Using 20 components in the atom index and the temporal mode and three components in the spatial mode the Tucker-3(20,20,3) model explains 80.2 % of the variance. Reducing the number of temporal and atom modes to 10 results in a data coverage of 60.6 %. A Tucker-3(4,4,3) model captures 36.9 % of the variance in the data.

The model constructed by the Protein Triads method is comprised of triads (figure 5.2). Combinations of a time mode, an atom mode and a spatial mode each have a corresponding value in the core array $\underline{\boldsymbol{G}}$. The data coverage for a triad is relative to the square root of its value for g. Orthogonal rotation of the core-array $\underline{\boldsymbol{G}}$ simplifies the model by maximizing the explained variance per triad [135]. Table 5.2 lists the results of a (10-10-3) Protein Triads analysis on the PYP data; the triads are ranked according to the explained variance. Comparing the first entries in table 5.2 shows that the first triad captures less data than the first eigenvector. Note that the various components in the Protein Triads model are sorted according to their appearance in a triad. Table 5.2 shows that the first atom mode is also present in the second triad, whereas their respective time and spatial modes differ. A similar pattern emerges in each of the models. The first two triads use the same atom mode in the (10,10,3), the (20,20,3) and the (4,4,3) model. Moreover, the explained variance covered by the first four triads is almost identical. The (4,4,3) model would be sufficient to describe the large concerted motions in this trajectory. Nevertheless, using the (10,10,3) model provides the means to distinguish between motions encompassing large parts of the protein and lesser motions.

Figure 5.5 (top) shows the atomic fluctuations resulting from the ED/PCA analysis on the PYP data. These are calculated from the scores in $\boldsymbol{T}$. Both eigenvectors show huge displacements for the N-terminal residues 1-3, and to a lesser extent for residues at positions 9 and 20 (ev1), 10 and 13 (ev2), and 48-49 and 98-99 (both). In the lower graph the atomic modes for the first three triads are displayed. The first atomic mode, present in triad 111 and 122, shows a large fluctuation for residues 1 to 3. Residues 9, 47 and 99 move in the opposite direction. In the third triad, 233, residues 6 and 20 move similarly, opposite to residue 9 and residues 69 and 112 in the PAS core. Similar flexible regions appear in the atomic profiles of ED/PCA and Protein Triads, located in the N-terminal cap and residues 47 and 99 in the PAS core. The pattern that emerges from the atomic profiles is however different. In ED/PCA, the N-terminal motion relates to different regions in the N-terminal cap (ev1: 9, 20, ev2: 10-13), but to similar regions in the chromophore binding pocket (residues 74 and 99). In contrast, the atomic profile from the Protein Triads analysis show that the N-terminal fluctuation occurs opposite to the motion exhibited by residues 9, 47 and 99. Additional motions in the N-terminal cap, region 6-20, do not appear to be related to the large fluctuations of the N-terminus. Interestingly, both methods pinpoint similar regions in the protein as highly flexible, regions that have emerged as functionally relevant (*i.e.* residues close to the chromophore and the N-terminal cap [57, 58, 60]). The Protein Triads analysis identified regions moving concertedly differently in comparison to ED/PCA.

To understand the cause underlying the differences in the results obtained from ED/PCA and Protein Triads, we decomposed the fluctuations of eigenvector 1 into its directional components x, y and z (see inset in figure 5.5). Component z exhibits a clear resemblance to the first atomic mode from the Protein Triads analysis, although some features also appear in the x and y components. In Protein Triads, these directional components are optimized to express the principal directions of internal motion in the protein. Figure 5.6 graphically displays the first and second spatial modes in PYP. To display optimally the fluctuations indicated in the Protein Triads analysis, those snaphots are selected that represent the highest and the lowest value for the relative fluctuations from the respective time profiles. Further investigation of the temporal profiles has little value. Since the parallel tempering trajectory contains contributions from various replicate MD simulations, it is not continuous. In both spatial modes, the residues highlighted close to the chromophore binding pocket move in opposite direction relative to the N-terminal motion. The orientation of the α-helix comprising residues 10-15 shifts in two directions, anti-correlated with the motion of the N-terminus. The shift of this helix is expressed in the third triad (see its atomic profile in figure 5.5). In conjunction with the rearrangements in the N-terminal domain, residues in the chromophore binding pocket shift position as well, as the fluctuations of residues 47 and 98 reflect. Although side chain interactions cause the changes in the chromophore binding pocket, backbone motions must accomodate the solvent exposure of the chromophore and Glu46. Glycine 47 is the hinge that allows solvation of Glu46, and Tyrosine 98 acts as a lid, opening up the chromophore binding pocket for exposure to solvent.

In agreement with previous analysis of the PYP parallel tempering trajectory and with experiments, the formation of the signaling state involves the unfolding of the chromophore binding pocket and structural rearrangements of the N-terminal domain. Here we have shown that these events occur in a concerted way: solvent exposure of Glu46 affects the conformation of the N-terminal cap.

Dynamics in different protein systems

In previous work we have explored the idea that similar structures induce a similar dynamic behaviour in a protein family. FixL, an oxygen sensor from the bacterium *Bradyrhizobium japonica* and HERG, the N-terminal domain of a potassium channel in *Homo sapiens* share a similar three dimensional fold. Using Molecular Dynamics trajectories of these two PAS folds we performed Protein Triads analyses to investigate whether these domains exhibit similar dynamics. The first five triads listed in table 5.3 cover approximately 45 % of the fluctuations in the simulation trajectories. In both systems, the time mode of the first triad appears also in later triads: For FixL triads 1, 2 and 4 share identical time profiles, and for HERG, the triads with identical temporal modes are 1, 3 and 4. This means that there is one large fluctuation underlying the variance in the systems, involving different atoms in different directions.

ED/PCA facilitates the comparison between different proteins, using separate analyses of the proteins of interest. Using the atomic variation, usually displayed as the root mean square fluctuation per atom, facilitates qualitative comparison, as Pandini *et al.* have shown for the PAS family [124]. Effective comparison requires knowledge on the structural similarities of the systems. Similarly, the atomic profiles resulting from Protein Triads analyses enable qualitative comparison. Figure 5.7 employs both atomic profiles and spatial modes to visualize the protein motions, and thus facilitates comparison between different protein systems. When exhibiting similar dynamics, the spatial modes point in identical directions. In figure 5.7 the proteins have similar orientations; the β-sheet is oriented such that the C-terminus points downward and all

FixL (Expl.var. = 73.8 %)

Component	Atom index	Time	Direction	Expl.var.(%)
1	1	1	1	22.7 (22.7)
2	2	1	2	9.4 (32.1)
3	3	2	2	5.8 (37.9)
4	4	1	3	4.1 (42.0)
5	5	3	3	4.0 (46.0)

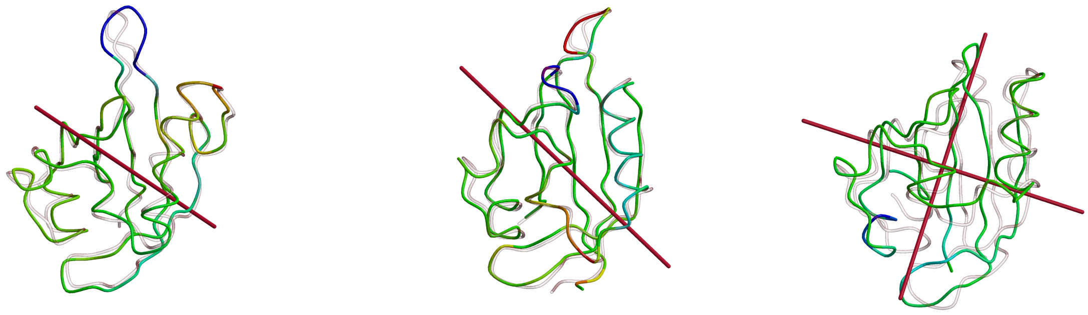
HERG (Expl.var. = 71.4 %)

Component	Atom index	Time	Direction	Expl.var.(%)
1	1	1	1	23.4 (23.4)
2	2	2	2	8.8 (32.2)
3	3	1	3	6.1 (38.3)
4	4	1	2	3.8 (42.1)
5	5	3	3	3.4 (45.5)

Table 5.3: **Details on the Protein Triads analyses of FixL and HERG.** The first five components are listed, sorted by explained variance. The explained variance (expl.var.) is given in percentages, with the summed variance in parentheses. For each triad (Protein Triads) the composition of modes from the atom index, the time and the direction is listed.

PAS α-helices are located in front of it. Highlighted in the protein structure are the regions that exhibit fluctuation, with red and blue parts moving in opposite directions along the directional mode. Green parts in the protein indicate regions with little fluctuation. These regions comprise residues in the central β-sheet in all three protein systems. In contrast, the location of the flexible regions in each PAS fold differs. Focusing in the PAS core of PYP, the residues exhibiting high fluctuations in PYP are centered around the chromophore, as discussed earlier. FixL contains a cofactor as well, located at the upper right side of the β-sheet in the orientation of figure 5.7(a). This is also the location of the highest fluctuations in FixL. No cofactor is present in the HERG PAS fold, but the protein domain contains a hydrophobic patch at the back of the β-sheet (in the orientation of figure 5.7(b)). The flexible regions in HERG comprise the loops connecting the strands in the β-sheet and the loop connecting helix $\alpha3$ and $\alpha4$ (nomenclature from Pandini *et al.* [124]). To summarize, the three protein systems share the central β-sheet as a stable structural element, but the systems differ in the location of more flexible features of the PAS folds.

Provided that the proteins are oriented similarly for ED/PCA, the overlap between the essential subspaces in the protein conformational space gives an indication of the similarity of the dynamics exhibited by the different proteins. The dimensions of the eigenvectors must be equal in size and are directly related to the number of atoms included in the model. As a consequence, only appropriate atoms are selected for the comparison of the dynamics of different proteins. This poses a severe restriction on the general applicability of ED/PCA in the comparison of dynamics in protein families. Protein Triads also facilitates direct quantitative comparison, not suffering from this restriction. Since the triads are orthogonal, the scalar product of the directional modes gives an indication of the similarity between the triads of different systems. Provided that the center of mass of the system is located at the origin, the value of the scalar product is a direct indication for the similarity in dynamics, since the length of the spatial component is



(a) FixL (b) HERG (c) PYP

Figure 5.7: **Graphical representation of the first triad in the FixL and HERG systems** The red axis indicates the first spatial mode as identified by Protein Triads analysis for (a) FixL, (b) HERG and (c) PYP. All proteins are oriented similarly with respect to the PAS β-sheet. The ribbon drawings represent the proteins in typical conformations that optimally display the amplitude of the motion, with the transparent and the colored coils respectively representing the highest and lowest values in the relative fluctuation in the time profiles of the first and second triad. The colors indicate the relative atomic fluctuation from blue (lowest value) to red (highest value) via green and yellow. Effectively, this means that the red and the blue parts move in an anti-correlated way. In the analysis of PYP, the first two components share the same atomic mode, hence two spatial modes are displayed. For clarity, the N-terminal cap of PYP is not shown.

equal to one. No fitting to an external reference frame is required for this comparison. Using the first triad of FixL and HERG, the scalar product is 0.55. This means that the angle between the two spatial modes is equal to 56^o (the cosine of the scalar product). Similarly, the angles between the first spatial modes of PYP and FixL, and of PYP and HERG are 54^o and 81^o. Such values indicate that these three domains do not compare well with regard to their dynamics.

Conclusion

The relevance of understanding the dynamic properties of proteins is being recognized more and more. Several techniques, including computer simulations, are available to provide data to support and generate hypotheses on protein dynamics, and link those to protein function. Unfortunately, these datasets are so complex that advanced statistical analysis methods are required to extract the relevant information. The method that is commonly used for such analyses is ED/PCA. This technique constructs a model for the data by summarizing the data in the temporal direction and in atomic fluctuations. When arranging the data from a molecular dynamics simulation as a three-way array, the application of three-way analysis techniques, such as Tucker-3 is possible. This method uses three sets of dynamically relevant components that describe (i) the relative fluctuation of the atoms, (ii) the time profile of a motion and (iii) the direction of a motion respectively. In this work we present the application of Tucker-3 analysis to protein dynamics: Protein Triads.

Starting with a simple system of eleven linked atoms undergoing an hinge bending motion, we show that Protein Triads performs equally well in capturing the variance in the data set as ED/PCA. Moreover, Protein Triads is able to separate the relevant motion into different model components, resulting in a better interpretable model. We then compared ED/PCA and Protein Triads analysis using a parallel tempering simulation of Photoactive Yellow Protein (PYP). Essential dynamics captures more variance in the first few model components than Protein Triads. Both methods indicate the N-terminal cap of PYP as the most flexible part in the protein. The model components obtained with Protein Triads analysis clearly distinguished motions in the N-terminal cap that are linked to the functional unfolding of the chromophore binding pocket, from flexibilities inherent to the N-terminal region.

PYP is a member of the PAS protein family, that is involved in signaling and protein-protein interactions. Using molecular dynamics simulations of the PAS folds in HERG, the N-terminal domain of a human potassium channel, and FixL, a bacterial oxygen sensor, we employ Protein Triads to compare the dynamics exhibited in the three protein systems in a qualitative and a quantitative way. Visualization of the model components facilitates qualitative comparison, provided that the proteins are oriented similarly. The scalar product of the directional modes is a direct indication for the similarity in dynamics between the datasets of the different proteins. This comparison shows that the PAS domains do not exhibit similar dynamics.

Chapter 6

Auto-inducer mediates TraR DNA binding through backbone fluctuations[*]

TraR is a quorum-sensing transcription factor from Agrobacterium tumefaciens that regulates genes required for conjugal plasmid transfer in presence of its auto-inducer 3-oxo-octanoyl-homoserine lactone (OOHL). It is functionally active as a dimer, protected from proteolysis by the auto-inducer. Comprising two domains, TraR contains an N-terminal PAS domain that binds OOHL and a C-terminal HTH-motif binding to DNA and is the shortest signaling pathway known that involves a PAS fold.

The crystal structure of TraR shows the protein binding DNA as a dimer, formed by two conformationally different monomers. Using this structural data we performed molecular dynamics simulations to investigate the dynamics involved in PAS mediated DNA binding. First we focus on the effect of the auto-inducer on fluctuations exhibited by the PAS domain, followed by a similar investigation for the full-length protein. Our simulations of the PAS domain revealed that the auto-inducer exhibited conformational heterogeneity, affecting surrounding residues in its binding pocket. The auto-inducer also affects the DNA binding domain in the elongated m2 conformation, whereas no clear difference emerges in the compact m1 conformation.

In absence of the auto-inducer, Tyr53 becomes solvent exposed. At elevated temperature, also Trp57 and Tyr61 are expelled from the protein interior. The simulation at high temperature also shows that the domains in the compact conformer move away from each other. These interdomain motions might eventually facilitate conversion of the compact monomer to the elongated conformation. Using Protein Triads to analyse the collective motions in the TraR monomers, we could correlate the enhanced fluctuations surrounding the auto-inducer to destabilization of the HTH-motif. Our observations lead to the postulation of a molecular mechanism for the DNA-binding of TraR.

Introduction

In the plant pathogen *Agrobacterium tumefaciens* quorum sensing regulates the replication and conjugal transfer of a plasmid that causes tumors in host plants. Similar to LuxR from *Vibrio fischerii* [138], the receptor protein TraR senses N-3-oxo-octanoyl-L-homoserine lactone (OOHL) [139] as auto-inducing molecule. In absence of the auto-inducer, TraR resides in the inner mem-

---

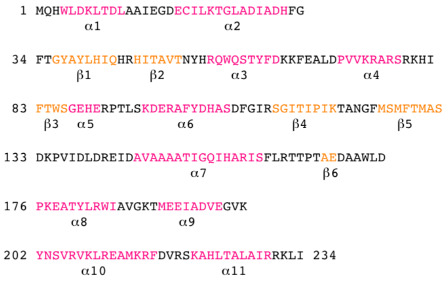
[*] This chapter is based on J. Vreede, K.J. Hellingwerf and W. Crielaard *Auto-inducer mediates TraR DNA binding through backbone fluctuations* Submitted to FEBS Letters

brane as a monomer. At sufficiently high concentrations of the auto-inducer OOHL, release of TraR into the cytoplasm occurs, where it forms dimers [140]. The dimeric complex of TraR with OOHL binds to nucleotide sequences called *tra*-boxes, activating transcription [141]. Several studies of the TraR protein and its homologues have indicated that this protein comprises two domains: the N-terminal domain binds one auto-inducer molecule, and the C-terminal domain binds DNA. The N-terminal domains mediate TraR dimerization, as evidenced by TraR inactivation by TrlR. The latter protein is similar to TraR, but lacks the last fifty amino acids. Similar to TraR, TrlR folds in presence of the auto-inducer, but does not bind DNA. In presence of TrlR, TraR activity drops, due to the formation of TraR:TrlR dimers [142].

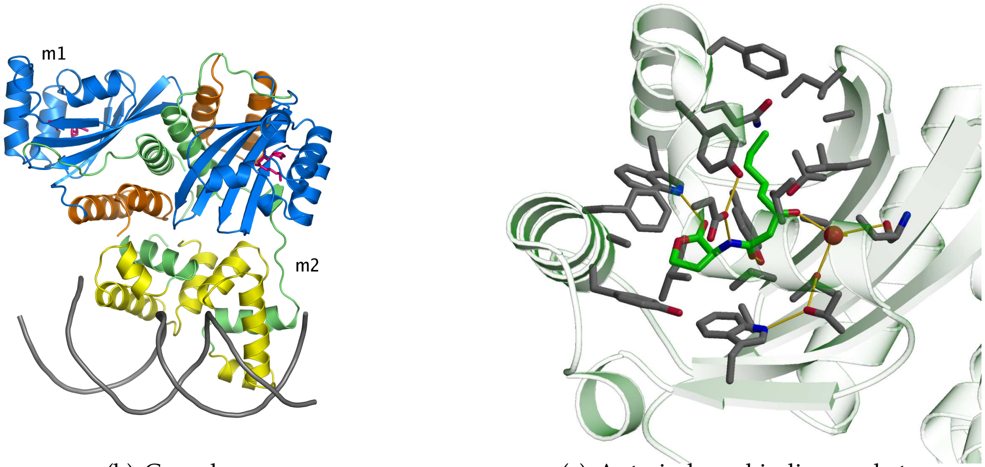
Two research groups simultaneously resolved the crystal structure of TraR. Both structures show TraR as a homodimer, complexed with DNA [143, 144]. The N-terminal domain belongs to the PAS family of signaling and dimerizing proteins, whereas the C-terminal domain shows a classical helix-turn-helix (HTH) motif for DNA binding. Both monomers bind the auto-inducer at the N-terminal domain, where it is embedded in the PAS fold. The dimer contains two regions where the two monomers interact, figure 6.1(b). As the first interface region, the two DNA binding domains form a symmetric dyad conformation. The two N-terminal domains also interact, and it is here that the two monomers exhibit conformational differences. The main deviation between the two conformations lies in the orientation of the N-terminal PAS domains towards each other. Central in the PAS fold is a five-stranded β-sheet, enclosing the auto-inducer. Three α-helices are located at the other side of the β-sheet, two N-terminal ($\alpha 1$, $\alpha 2$) and one C-terminal ($\alpha 7$) to the PAS domain (see figure 6.1(a) for nomenclature). These form the interface between the PAS fold and the HTH motif in the m1 monomer. In the m2 monomer, these helices interface with the other monomer. As a consequence, this monomer is more elongated, with only a few electrostatic side chain interactions between the PAS domain and the HTH motif. Also, the PAS β-sheet is extended with an additional strand by the linker region connecting the two domains.

Recently, the solution structure of the N-terminal domain of SdiA, a quorum sensor from *Escherichia coli*, became available [145], in complex with N-octanoyl-L-homoserine. The structure exhibits a PAS fold, closely resembling the N-terminal domain of TraR. Although the auto-inducer is deeply embedded within the PAS fold, the SdiA structural ensemble diplays conformational heterogeneity within the binding pocket of the auto-inducer. The SdiA solution structure, in combination with the TraR crystal structures show that the bacterial quorum sensors for acyl lactone serines comprise the shortest signaling pathway involving a PAS domain.

The current view on the role of the auto-inducer is that it acts as a folding switch. Whether it is also involved in mediating the DNA binding activity of TraR is unknown. Using molecular dynamics simulations we investigate the effect of OOHL on the fluctuations exhibited by TraR. First the focus lies on the effect of the auto-inducer on fluctuations within the PAS domain, followed by a similar investigation for the full-length protein. For the latter, we also performed Protein Triads [146] analyses to visualize the collective motions within the TraR monomers. Our observations enabled us to postulate a molecular mechanism for the DNA-binding of TraR. Finally, sampling conformations at elevated temperature allowed identification of residues possibly signaling for proteolysis.


1 MQHWLDKLTDLAAIEGDECILKTGLADIADHFG
α1
α2
34 FTGYAYLHIQHRHITAVTNYHRQWQSTYFDKKFEALDPVVKRARSRKHI
β1
β2
α3
α4
83 FTWSGEHERPTLSKDERAFYDHASDFGIRSGITIPIKTANGFMSMFTMAS
β3
α5
α6
β4
β5
133 DKPVIDLDREIDAVAAAATIGQIHARISFLRTTPTAEDAAWLD
α7
β6
176 PKEATYLRWIAVGKTMEEIADVEGVK
α8
α9
202 YNSVRVKLREAMKRFDVRSKAHLTALAIRRKLI 234
α10
α11

(a) Sequence



(b) Complex

(c) Auto-inducer binding pocket

Figure 6.1: **Sequence and structure of TraR** (a) The amino acid sequence of TraR, with secondary structure elements indicated in pink for α-helices and orange for β-strands. (b) The crystallized TraR-DNA complex in ribbon representation: orange, N-terminal to the PAS fold; blue, the PAS fold; green, linker region; yellow, the HTH motif; grey; DNA. The pink stick model represents the auto-inducer. (c) The binding pocket of the autoinducer in stick models. Carbon atoms of the auto-inducer are highlighted in green. The yellow lines represent hydrogen bonds

Methods

Using molecular dynamics simulations, we investigated the dynamic behaviour of TraR in presence and absence of the auto-inducer. Structural data on TraR shows a complex containing two conformationally different monomers bound to a stretch of DNA. Both monomers contain an N-terminal PAS fold as an input sensor domain and a C-terminal HTH-motif as a DNA binding output domain. First we focused on the effect of the signaling trigger on the dynamics of the PAS domain, followed by sampling of the conformational space of the full-length protein, using both conformers.

Figure 6.1(b) was created with Pymol [147], all other protein drawings were created with Molscript [136] and Raster3D [137].

Molecular Dynamics simulations

As a starting point, we chose the crystal structure at 1.66 Å resolution (pdb-code 1L3L [144]). This structure comprises a dimer that binds one pieces of DNA. Decomposing the dimer reveals its monomers in compact and more elongated conformations. The more elongated monomer lacked coordinates for a flexible loop between the two domains (residues 163-171). Using the structure at 3.0 Å resolution (pdb-code 1HOM [143]), preliminary models for these disordered regions were obtained. Similarly, surface residues lacking side chain atoms were completed.

The completed structures served as input for several molecular dynamics simulations: both monomers, the sensor domains of the monomers, with and without the autoinducers. For the latter, omission of the coordinates of the autoinducer sufficed. We did not attempt to perform simulations of the full dimer system due to restrictions in size. After addition of hydrogen atoms at polar and aromatic positions the molecules were solvated in a cubic periodic box, with at least 1 nm between the protein and the box boundaries. SPC water molecules [90] were added to the systems in addition to the crystal waters. Removal of waters overlapping with or located within hydrophobic cavities in the protein preceded the replacement of water molecules at the most electropositive locations with chlorine ions. Prepared as such, all systems were energy minimized (200 steps of conjugated gradient), followed by the system equilibration to dissipate excess energy. Water and hydrogen positions were equilibrated for 10 ps and full-system equilibration lasted 1 ns in the NpT ensemble. Prepared as such, the systems were used as input for multiple runs (of 10 ns per run) of molecular dynamics.

Atomic interactions were described by the GROMOS united atom force field [133, 134] in combination with the SPC water model [90]. Parameters for the autoinducer were obtained from the PRODRG server [148] and adapted to the GROMOS force field. Van der Waals interactions were treated with a cut-off at 1.4 nm, and PME handled the long range electrostatics. The use of constraints, LINCS for solute interactions [95] and SETTLE for water interactions [96], allowed for a time step of 2 fs. System temperature was kept constant using the Berendsen thermostat [97], and constant pressure was achieved using the Berendsen barostat [97]. The resulting trajectories served as input for several analyses using the gromacs tools. Analyses comprised calculation of root mean square fluctuations in the positions of C_α atoms and analysis of the hydrogen bonding network in the auto-inducer binding pocket.

Protein Triad analysis

In a trajectory resulting from an MD simulation, many motions are sampled. The Protein Triad analysis method [146] decomposes a trajectory into components that aim to explain the variance exhibited in the simulation. These components, referred to as triads, contain three modes that describe the motions dominating the variance in the trajectory. Atom or residue index, time and direction make up the three modes in a triad. We applied Protein Triad analysis to the simulations of the full-length TraR conformations. Prior to the analysis, the MD trajectories were aligned with their starting positions, using the C_α positions to minimize the root mean square deviation between the conformations. For the extraction of the collective motions we used MATLAB in combination with the Nway toolbox [135], with 10 components for the atomic and temporal modes and 3 components for the directional modes.

Results and discussion

The quorum sensor TraR from *Agrobacterium tumefaciens* acts as a transcription factor, regulating genes required for conjugal plasmid transfer in the presence of its auto-inducer 3-oxo-octanoyl-homoserine lactone (OOHL). TraR contains an N-terminal PAS domain that binds OOHL and a C-terminal HTH-motif interacting with DNA, and thus comprises the shortest signaling pathway involving a PAS fold as input domain. The currently established role of the auto-inducer is that of the folding switch, protecting TraR from proteolysis. After folding, TraR forms a dimer, as indicated by several experiments and further substantiated with crystallographic structural data. Using the latter, we performed molecular dynamics simulations on the PAS domain and the full-length TraR to investigate the effect of the auto-inducer on the dynamic behaviour of the protein. Also, we performed simulations at 400 K to investigate unfolding events, allowing the identification of proteolysis signals. The total simulation time is 40 ns for the TraR PAS domain is 40 ns and 60 ns for the full-length TraR.

Fluctuations in the PAS domain

As figure 6.1 displays, the crystal structure of TraR shows a TraR homodimer in complex with DNA. Although the two monomers are identical in chemical composition, their conformations differ, distinguishing a compact (m1) and an elongated monomer (m2). The PAS domain however is similar in both conformations. Inside the binding pocket, buried within the PAS fold, the auto-inducer interacts with the protein through hydrogen bonds with Trp57, Asp70 and Tyr53, a water mediated hydrogen bond with Thr129, while its aliphatic chain sticks into a hydrophobic pocket. The water-mediated hydrogen bond involves a water molecule buried in the protein core and is further stabilized by hydrogen bonds to the backbone of the PAS-β-sheet (Ala38, Thr115, Met127). The auto-inducer is shielded from water interactions by Gln58 and Tyr102.

To investigate the effect of the auto-inducer on the dynamics of the TraR PAS domain, we used residues 1-162 to set up molecular dynamics simulations in presence and in absence of the auto-inducer. This stretch of amino acids contains the three interfacing helices (α1, α2 and α7)in addition to the PAS fold. Figure 6.2 shows the root mean square fluctuations per residue obtained from simulations of the PAS domains in the two TraR monomers. The upper panel

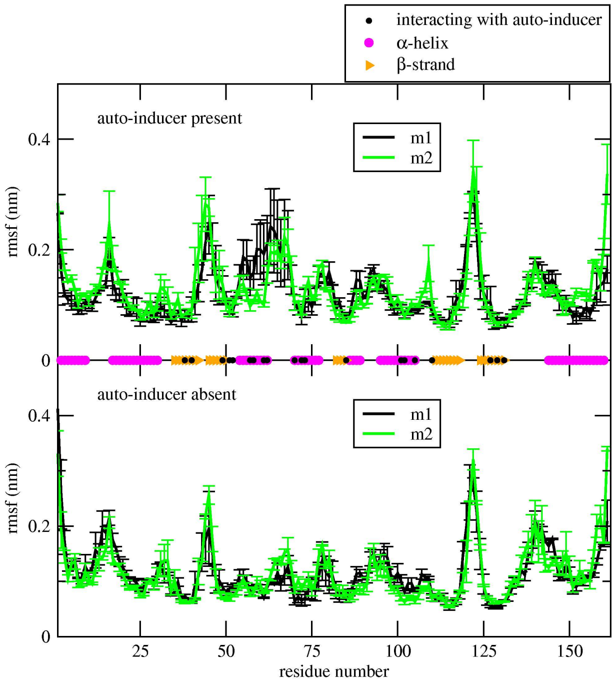


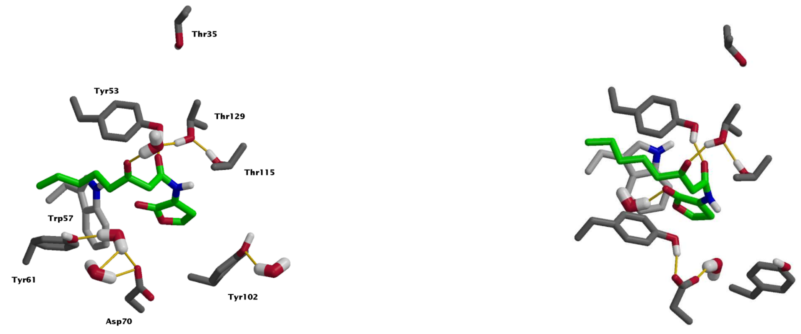
Figure 6.2: **Root mean square fluctuations in TraR-PAS** Deviations in atomic displacement are averaged in nm for each residue in residues 1-162 of TraR. Averaging over two molecular dynamics runs, the error bars indicate the drift in sampling. Secondary structure is highlighted at the horizontal axis with magenta circles indicating α-helices and orange triangles indicating β-strands. The black dots indicate residues within a distance of 0.6 nm of the auto-inducer.

displays the PAS folds including the auto-inducer and the lower panel shows the fluctuations in absence of the auto-inducer. In both graphs, the largest fluctuations, with a value of 0.3 nm occur in the loop region around residue Gly123, connecting strands $\beta4$ and $\beta5$. Mutation of this residue has affected the transcriptonal activity of TraR, rather than its DNA binding affinity [149]. This study implicates Gly123 in modulating the binding of RNA polymerase in vivo. Changing its conformational properties would severely affect the flexibility of TraR.

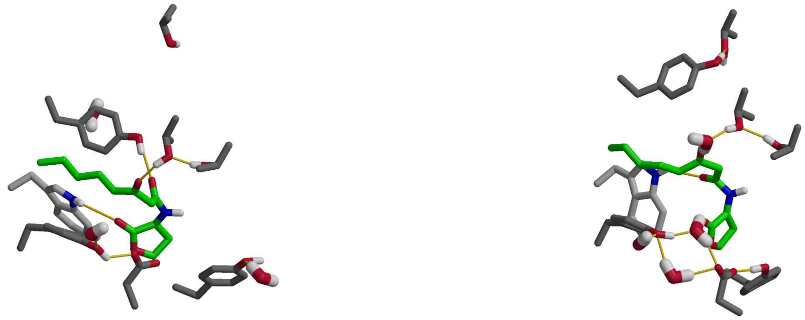
The rmsf profile matches the secondary structure of the TraR input domain: peaks in the profile correspond to loops, whereas the β-strand regions move the least. As expected, the residues that interact with the auto-inducer show different behaviour. Residues 55-63, comprising the helix $\alpha3$ of the PAS fold show more displacements in presence of the auto-inducer. Focusing on the binding pocket of the auto-inducer, figures 6.3 and 6.4 shows typical snapshots taken during the MD simulations of the PAS domain. All systems containing the auto-inducer started with an intact binding pocket with the hydrogen bonds as highlighted in figure 6.1(c) intact. During the molecular dynamics runs, the hydrogen bonding network undergoes a number of rearrangements. Asp70 acts as an hydrogen bond acceptor for the NH group of the auto-inducer and for the OH group of Tyr61. Since these residues are solvent accessible, water molecules start to interfere with the latter hydrogen bond, disrupting interactions between Asp70 and the auto-inducer. The auto-inducer has now more freedom of motion, shifting its hydrogen bond to Trp57 and ultimately break it. Another consequence of the increased flexibility of the auto-inducer is that the water incorporated in the binding pocket is expelled. When the auto-inducer is absent, water molecules enter the hydrophilic part of the auto-inducer binding pocket. The internal water molecule stays hydrogen bonded to Thr115 and backbone atoms. Water molecules in close contact with Asp70 and Tyr61 form an interacting chain connecting the internal water to the bulk. When Tyr102 participates in the hydrogen bond network around Asp70, it blocks water molecules from entering the protein interior. Early in the simulations, Tyr53 becomes exposed to solvent, and one run shows it interacting with Thr35.

Our simulations of the TraR PAS domain show increased flexibility around the binding pocket of the auto-inducer when the compound is part of the system. In contrast, the crystal structure of TraR shows the auto-inducer embedded within the crystal structure in a single conformation. Recently, the NMR structure of the PAS domain in the *E. coli* quorum sensing protein SdiA became available. This protein acts as a folding switch in the presence of N-octanoyl-L-homoserine lactone, similar to TraR. Its structure shows the auto-inducer deeply embedded in a PAS fold, where it exhibits conformational heterogeneity. Also Asp80, homologous to Asp70 in TraR, assumes various conformations [145]. These observations are in agreement with our simulations of the PAS fold of TraR. Further proof of the flexibility of the auto-inducer binding pocket is provided by a study investigating the kinetics of TraR binding different auto-inducers in vivo. The authors show that TraR is able to bind auto-inducers with a smaller aliphatic chain (6 carbon atoms instead of 8 in the endogenous auto-inducer) [150]. SdiA also shows this behaviour: it folds into a soluble globular protein in the presence of different auto-inducers [145].

As shown in our simulations, the auto-inducer has inherent flexible properties that affect residues in its proximity, in agreement with NMR data. Also, residues interacting with water molecules induce conformational freedom in the auto-inducer. These protein-solvent interactions originate from the solvation of the crystal structure in a water box.

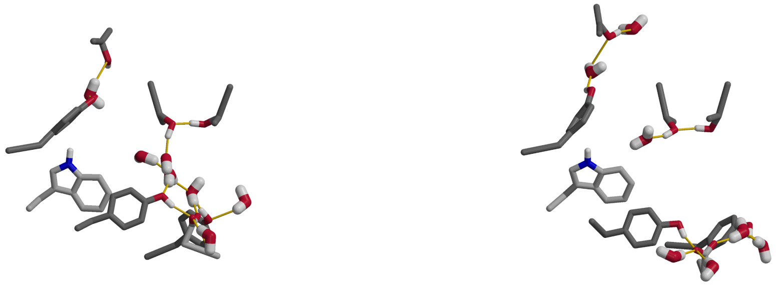


(a) auto-inducer present m1 (b) auto-inducer present m1



(c) auto-inducer present m2 (d) auto-inducer present m2

Figure 6.3: **Snapshots of the auto-inducer binding pocket from TraR-PAS simulations.** The snapshots display typical configurations of the auto-inducer binding pocket in stick configuration. Carbon atoms of the auto-inducer are highlighted in green. The yellow lines indicate hydrogen bonds. Residue names are indicated in (a), all configurations are oriented similarly.



(a) auto-inducer absent m1

(b) auto-inducer absent m2

Figure 6.4: **Snapshots of the auto-inducer binding pocket in absence from the auto-inducer.** The snapshots display typical configurations of the auto-inducer binding pocket in absence of the auto-inducer in stick configuration. The yellow lines indicate hydrogen bonds. Residue names are indicated in figure 6.3(a), all configurations are oriented similarly.

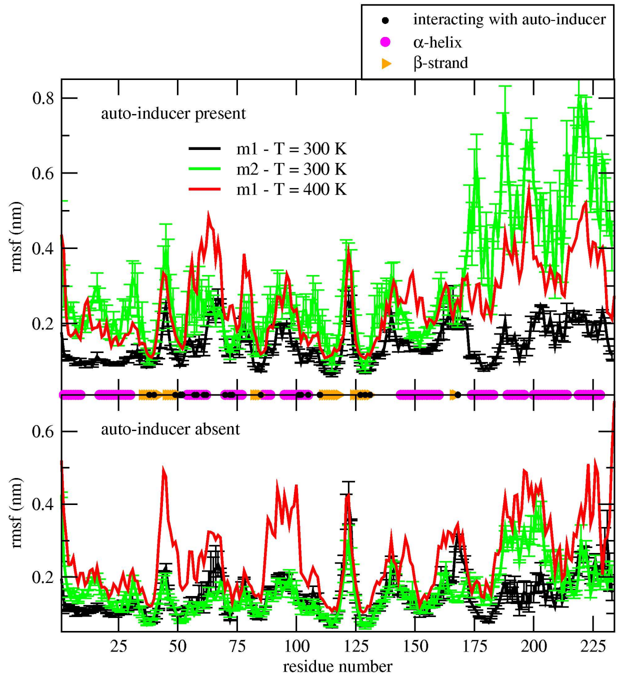


Figure 6.5: **Fluctuations in full-length TraR** Deviations in atomic displacement are averaged in nm for each residue in TraR. Averaging over two molecular dynamics runs, the error bars indicate the drift in sampling. Secondary structure is highlighted at the horizontal axis with magenta circles indicating α-helices and orange triangles indicating β-strands. The black dots indicate residues within a distance of 0.6 nm of the auto-inducer.

The buried water molecule found in the crystal structure of TraR might be stable when the protein is constrained in a crystalline lattice, but is expelled and/or replaced by other water molecules in a solution environment.

Mutagenesis experiments investigating the auto-inducer affinity of TraR show that changing Thr115 and Thr129 to hydrophobic groups affects the ability of TraR to bind the auto-inducer [151]. In our simulations we show that these residues are an integral part of a hydrophilic, hydrated cavity in TraR, in contact with bulk water. Moreover, solvent accessibility of the auto-inducer binding pocket may explain the observation that TraR has a lower affinity for homoserine-L-lactones lacking the 3-oxo group. In absence of this group, activation required the continuous presence of the external signal, in high concentrations [150]. The presence of any auto-inducer suffices to activate the folding of TraR. Due to the solvent accessibility of the binding pocket of TraR, and the absence of the oxo-group, reducing the number of hydrogen bonds between auto-inducer and protein, the auto-inducer may diffuse out of the binding pocket. Unless present in high quantities, thus increasing the probability of the auto-inducer being inside the protein, transcriptional activity is reduced.

Interactions between the PAS fold and the HTH motif

The MD simulations on the PAS domain of TraR have shown that the auto-inducer enhances fluctuations in its direct environment. To investigate whether the auto-inducer also influences displacements farther away, *i.e.* in the HTH motif, we performed simulations of full-length TraR, starting from both monomers available from the crystal structure. We did not perform simulations of the full-length dimer, due to the large size of the system. Figure 6.5 shows the fluctuations in the two TraR monomers as a function of the residue number. When the auto-inducer is present in the simulations, the two monomers exhibit different fluctuations. Overall, larger fluctuations occur for conformation m2 in comparison to m1. The HTH motif shows the most striking differences; in m1 the domain doesn't have displacements higher than 0.2 nm, whereas the fluctuations exceed 0.6 nm for m2. To a lesser extend, the two N-terminal α-helices exhibit different fluctuations. In m1 these helices are restricted in their conformational freedom, since they are part of the interface between the PAS domain and the HTH motif. The m2 conformation allows more flexibility for these helices, as reflected by their larger fluctuations. Stable structural elements in all simulations are the PAS β-sheet and the helix in between the PAS domain and the HTH motif.

Regardless of the presence of the auto-inducer, the PAS fold is a stable protein configuration for the duration of the molecular dynamics sampling. We performed simulations of both TraR monomers to investigate whether this also holds for the full-length protein. The lower panel of figure 6.5 shows the fluctuations in both monomers. In m1 little differences appear when comparing the simulation with and without auto-inducer, apart from the differences in the binding pocket. The absence of the auto-inducer has a clear effect on the HTH motif of m2 as well as on the two N-terminal helices: the fluctuations have decreased significantly.

To visualize the large collective motions exhibited by the TraR conformations, we performed Protein Triads analyses on the trajectories of the full-length protein simulations. The Protein Triads analysis method extracts the collective motions from a set of conformations [146], such as generated during molecular dynamics sampling. The collective motions are represented as triads, see table 6.1, that have a residue component, a time component and a directional component. Applied to the TraR full length simulations, Protein Triads, using 10 components for the atom component and the time component, captures more variance for m2 (90%) than for m1 (75%). In all analyses the first two triads share a time mode, and the first triad (as sorted by explained variance) explains 8-18 % more than the second. Only the analysis of the m2 trajectory in presence of the auto-inducer contains a significant contribution of another time mode.

TraR-full monomer 1

Auto-inducer present (Expl.var. = 70.6 %)				Auto-inducer absent (Expl.var. = 77.9 %)			
Res. index	Time	Direction	Expl.var.(%)	Res. index	Time	Direction	Expl.var.(%)
1	1	1	18.5	1	1	1	30.0
2	1	2	10.0	2	1	2	12.0
3	1	3	7.8	3	1	3	9.1

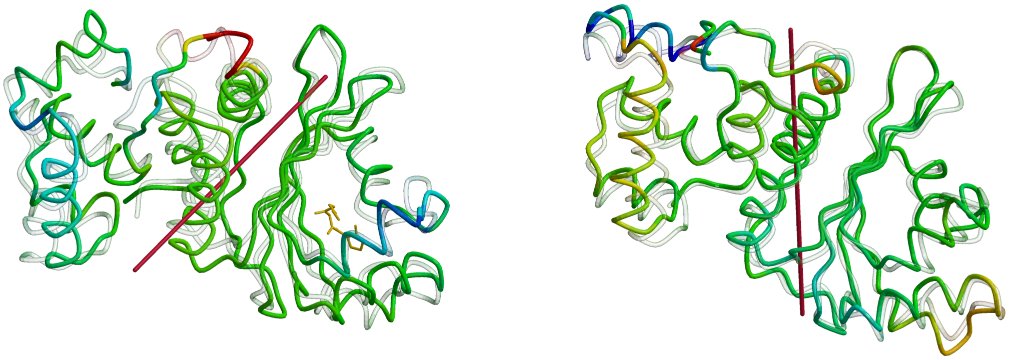
TraR-full monomer 2

Auto-inducer present (Expl.var. = 93.6 %)				Auto-inducer absent (Expl.var. = 89.6 %)			
Res. index	Time	Direction	Expl.var.(%)	Res. index	Time	Direction	Expl.var.(%)
1	1	1	25.3	1	1	1	34.2
2	1	2	17.4	2	1	2	25.7
3	2	1	14.8	3	2	1	8.4
4	2	3	12.8				

Table 6.1: **Protein Triads analysis of the MD simulations of the TraR monomers m1 and m2 in presence and absence of the auto-inducer.** The explained variance (Expl. var.) of the complete Protein Triads analysis is given between brackets, followed by a list of the triads. Each triad contains a residue component (Res. index), a time component and a directional component. The table lists the composition of the triads, as well as the explained variance. Triads are only listed if their explained variance is higher than 10 % or if they share a mode with preceding triads.

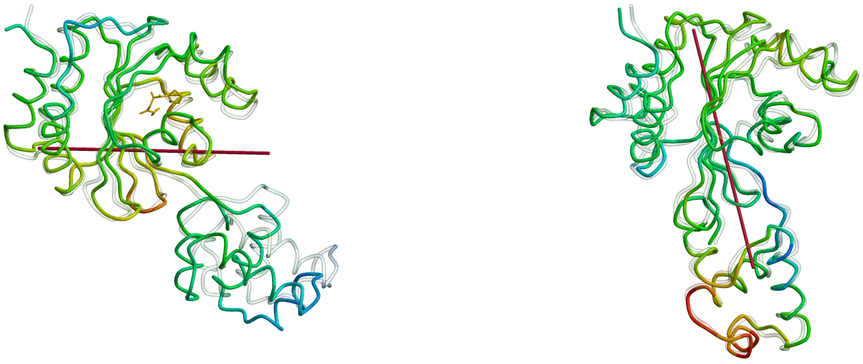
Figure 6.6 visualizes the first triads obtained from the simulations of the full-length TraR conformations. The compact m1 conformation, in presence of the auto-inducer (figure 6.6(a)), shows high flexibility for the linker region (red), that is oppositely correlated to motions of helices α3 close to the auto-inducer and α10, the recognition helix in the HTH motif (blue). When the auto-inducer is absent (figure 6.6(b)), anti-correlated fluctuations of helices α10 (yellow) and α11 (blue) dominate the variance in the simulation, in conjunction with helices α5 and α6 and the flexible linker region. In the more elongated m2 conformation, the Protein Triads visualizations clearly show that the auto-inducer affects the motions in the protein. In absence of the auto-inducer, the fluctuations are mainly located in the HTH motif. When the auto-inducer is present (figure 6.6(d), helix α6 and the loops connecting β1 to β2, and β4 to β5 show larger fluctuations, as well as the additional strand β6 (yellow). Helix α6 moves opposite to these elements, in concert with helix α9 in the HTH-motif (blue). Moreover, the HTH motif moves relative to the PAS fold. Since this does not occur in absence of the auto-inducer, the domain movement is a consequence of the higher flexibility induced by the auto-inducer.

Our observations show that m1, as a compact conformation, is stable and exhibits largest fluctuations in its flexible linker region. The m2 conformer exhibits more flexibility, and more importantly, it shows significantly different behaviour in presence and absence of the auto-



(a) ai present m1

(b) ai absent m1



(c) ai present m2

(d) ai absent m2

Figure 6.6: **Visualisations of the first triads in the TraR simulations** The coil structures represent the highest (transparent) and the lowest value along the time mode. The coloring displays the relative atomic fluctuation per residue, with blue for the lowest value via green and yellow to red for the highest value. Shown as a red axis, the spatial mode indicates the direction of the motion described by the first triad. Yellow stick representations display the auto-inducer, when present. All conformations are oriented such that the PAS β-sheet is perpendicular to the viewing plane.

inducer. In the latter system, the HTH motif and the PAS fold move closer to each other in absence of the auto-inducer. When present, the auto-inducer affects surrounding residues, including Gly123, preventing the HTH to move toward the PAS fold. Instead, the HTH motif assumes various different orientations with respect to the PAS fold. The inactivation of TraR by TrlR provides further evidence for the role of the PAS domains in forming the dimer. TrlR is a TraR-like protein lacking fifty C-terminal residues. Inactivation of TraR proceeds through the formation of a TraR-TrlR dimer that is unable to bind DNA or activate transcription [142].

Destabilization of the compact monomer

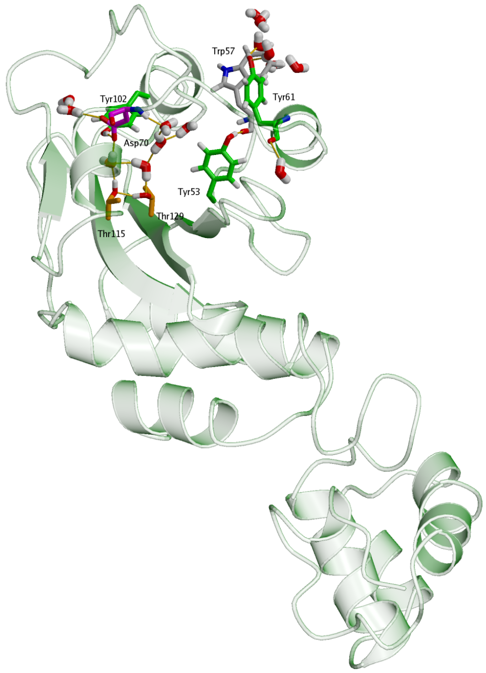


Figure 6.7: **Snapshot of m1 of the simulation at 400 K** Cartoon representation of TraR with residues Trp57 highlighted in grey, residues Tyr53, Tyr61 and Tyr102 highlighted in green, residues Thr115 and Thr129 highlighted in orange and residue Asp70 highlighted in magenta stick models. Water molecules within a distance of 4 Å are displayed as stick models. The yellow lines represent hydrogen bonds. The backbone atoms of the highlighted residues are not shown, except for Trp57, Tyr61 and Asp70.

In absence of the auto-inducer, proteolysis of TraR occurs, when it is present in the cytoplasm. The assumption is that the protein is unfolded, with residues exposed to solvent that signal for proteolysis. Although no unfolding events occurred, Tyr53 quickly rotates out of the auto-inducer binding pocket toward solvent interactions, in the simulations without auto-inducer. This (partial) solvent exposure might be a first step toward the unfolding / activation of proteolysis of TraR. Increasing the temperature in a simulation, while keeping the volume fixed allows the crossing of energy barriers, speeding up the sampling of the conformational space. Although such simulations do not represent equilibrium properties, they indicate stable and labile regions in a protein. We performed molecular dynamics simulations at 400 K of the m1 conformation in presence and absence of the auto-inducer, and displayed the protein fluctuations in figure 6.5, red lines. When the auto-inducer is present, the fluctuation profile of m1 at 400 K has a closer resemblance to conformation m2 at 300 K than to conformation m1 at 300 K, in particular on the DNA binding region. In addition, and not observed in the 300 K simulations, the 400 K simulation shows large fluctuations for residues 64-69. Comparing the simulations at 300 K and at 400 K in absence of the auto-inducer shows that the latter has significantly increased fluctuations for residues in the auto-inducer binding pocket: 45-47, 57, 64-69 and residues in helices $\alpha4$, $\alpha5$ and $\alpha6$. Within the HTH-motif residues 221-229 and to a lesser extent residues 188-223 exhibit larger fluctuations. Comparing the simulations at 400 K shows that residues 64-69 exhibit enhanced fluctuations in presence of the auto-inducer, whereas its absence destabilizes residues 45-47 and helices $\alpha5$ and $\alpha6$. These results show that the auto-inducer influences the behaviour of surrounding residues.

Focusing on the fluctuations in the auto-inducer binding pocket, the simulation in absence of OOHL starts with a hydrated auto-inducer binding pocket, similar to the water channel in figure 6.4(a) and with all residues originally involved in hydrogen bonds inside the protein. During the high temperature sampling, Trp57 and Tyr61 become solvent exposed, shifting the orientation of helix $\alpha6$. In contrast, Trp57 remains buried in the protein core in presence of the auto-inducer. Fusion proteins between the N-terminus of TraR and the N-terminus of a acetyltransferase make TraR more resistant against proteolysis, in absence of the auto-inducer. These results suggest that N-terminal residues of TraR signal for proteolysis. Also, although the fusion proteins protect TraR from proteolysis, the DNA-binding activity of the fusion proteins strongly increased in the presence of the auto-inducer, indicating that its function comprises more than shielding TraR from proteolysis [152]. The 400 K simulation of TraR conformer m1 indicates Trp57 and Tyr61 as a possible signal for proteolysis, rather than N-terminal residues.

The simulations at elevated temperature also show fluctuations in the HTH-motif: helix $\alpha11$ unfolds, while helices $\alpha9$ and $\alpha10$ in the HTH motif rotate with respect to each other. Most interestingly however, the DNA binding domain moves away from the PAS domain, breaking a salt bridge between Asp10 and Arg183. In presence of the auto-inducer, this salt bridge forms again during the simulation, within 5 ns after its rupture. The simulations at 300 K showed that the auto-inducer enhances fluctuations in its direct environment, and also affects the relative orientation of the HTH motif toward the PAS fold in the elongated m2 conformation. No clear correlation appeared for the compact m1 conformation between dynamic behaviour of the protein and the presence of the auto-inducer. Assuming that m2 is a functional conformation, the TraR dimer would be more effective in binding DNA with two loose HTH motifs. The high temperature simulation of m1 displays an interdomain motion that might facilitate conversion of m1 into m2.

Combining the observations described here caused us to postulate the following mechanism for TraR DNA binding. The PAS domain in both monomers is stable and is involved in the dimer formation of TraR [142]. With the interacting PAS domains as a stable scaffold, the auto-inducers enhance fluctuations in their proximity. This prevents the HTH motifs from interfacing with the

PAS domains, resulting in a dimer of two elongated m2 conformations. As a consequence, the two HTH motifs dangle and search their surroundings for DNA. Once in contact with DNA, both HTH motifs interact to form a symmetric dyad interface. As a result, one monomer shifts into the compact m1 conformation, accomodating the DNA.

Conclusion

TraR is a quorum-sensing transcription factor from *Agrobacterium tumefaciens* and regulates genes required for conjugal plasmid transfer in presence of its auto-inducer 3-oxo-octanoyl-homoserine lactone (OOHL). It is functionally active as a dimer, protected from proteolysis by the auto-inducer. Comprising two domains, TraR contains an N-terminal PAS domain that binds OOHL and a C-terminal HTH-motif binding to DNA. Structural data of TraR shows the protein binding DNA as a dimer, comprising two conformationally different monomers. One monomer is compact, whereas the other has a more elongated configuration.

The current view on the role of the auto-inducer is that it acts as a folding switch. Whether it is also involved in mediating the DNA binding activity of TraR is unknown. In this work we investigated the effect of the auto-inducer on the fluctuational behaviour of the protein. Using both monomer conformations we performed molecular dynamics (MD) simulations of TraR.

Our simulations revealed that the auto-inducer and its close environment show conformational heterogeneity, involving several water molecules. Moreover, it is an interplay between the auto-inducer and water molecules in its proximity that induces the conformational variety. This heterogeneity is in agreement with the NMR structural ensemble of SdiA, a related protein from *Escherichia coli*. The auto-inducer causes the heterogeneity: In its absence, Tyr53 becomes solvent exposed and the protein is less flexible. Moreover, the auto-inducer also affects residues not in its direct environment. The DNA binding domain shows significantly enhanced fluctuations in presence of OOHL in the elongated m2 conformation. In other words, the presence of the auto-inducer in the PAS domain induces fluctuations in the HTH motif that are instrumental for DNA binding.

In the MD simulation at elevated temperature, also Trp57 and Tyr61 are expelled from the protein interior, in absence of OOHL. These residues might be involved in signaling for proteolysis. The simulation at high temperature also shows that the domains in the compact m1 conformer move away from each other. These interdomain motions might eventually facilitate conversion of the compact monomer to the elongated conformation.

Summarizing these observations results in a mechanism for the DNA binding of TraR and possibly its sensitivity to proteolysis. The auto-inducers in the TraR dimer induce increased fluctuations in their close surroundings. As a consequence, the HTH motifs dangle from the PAS scaffold, searching for DNA. Binding of DNA induces the formation of an interface between the two HTH motifs, and is accomodated by the formation of the more compact m1 conformer. The conformational change, in absence of OOHL and at high temperature, leading to the solvent exposure of Trp57 and Tyr61 is indicative for protein unfolding and hence sensitivity for proteolytic enzymes.

Chapter 7

Discussion

The PAS domain sequence was first recognized in transcription factors of *Drosophila* as two imperfect repeats [23]. Later on, these repeats have been identified as separate PAS domains [14], sometimes containing a additional PAC sequence. A recent computational study used an automated modeling procedure to produce 3D structural models of all known PAS sequences and showed that the PAC sequence is part of the PAS fold, but exhibits low sequence similarity [153]. Currently, a search in the Protein family database with PAS as a keyword results in finding the PAS superfamily, also known as clan. This illustrates the expansion of knowledge on the PAS protein family since its discovery in 1997. This chapter provides a context in the expanding knowledge on PAS domains.

At the start of this work, structural data on four PAS domains was available: Photoactive Yellow Protein, a blue light receptor from the bacterium *Halorhodospira halophila*, the sensor domain of FixL, an oxygen sensor originating from rhizobia, LOV, a light receptor found in ferns and HERG, the N-terminal domain of a human potassium channel. These PAS domains form the basis of the research carried out in chapters 2-5. More structures meanwhile became available, however, including the one of the quorum sensor TraR from *Agrobacterium tumefaciens*. This protein is the main topic of Chapter 6. Currently, structural data of thirteen PAS domains has become available. Details are listed in table 7.1.

All structures listed in table 7.1 have a central five- or six-stranded β-sheet. Using this as a reference, all structures are oriented similarly for display, see figure 7.1(a). The prototype of the family, PYP, is displayed as a cartoon representation in figure 7.1(b), highlighting the structural features of the PAS family. Central in the structures is the antiparallel β-sheet with three or four helices located at its concave side. The β-sheet is a very stable structural element; molecular dynamics simulations of PYP at 500 K, in a constrained volume, show that this sheet remains intact, even under these extreme conditions*. This simulation also showed that the helices lying adjacent to the β-sheet are less stable (see Chapter 3 for more details). The orientation of these helices varies in the different PAS domains, to accomodate co-factors and/or interaction with other protein domains.

The currently known PAS domains bind a variety of cofactors, covalently, via (water-mediated) hydrogen bonds or via hydrophobic interactions. Figure 7.1(b) shows the location of all known

---

*Note that this is not a physical property, but a feature of the force field used for the simulation.

Protein	Function	Organism	Method	PDB code	references
FixL	oxygen sensor	*Rhizobia meliloti*	X-ray	1D06	[154]
		Bradyrhizobia japonica	X-ray	1DRM	[36]
DOS	oxygen sensor	*Escherichia coli*	X-ray	1V9Z	[155,156]
PYP	photoreceptor	*Halorhodospira halophila*	X-ray	1NWZ	[157]
	photoreceptor	*Rhodospirillum centenum*	X-ray	1MZU	[158]
LOV2	photoreceptor	*Adiantum capillus-veneris*	X-ray	1G28	[37]
LOV1	photoreceptor	*Chlamydomonas reinhardtii*	X-ray	1N9L	[159]
CitA	citrate sensor	*Klebsiella pneumonia*	X-ray	1POZ	[122]
TraR	quorum sensor	*Agrobacterium tumefaciens*	X-ray	1LL3, 1H0M	[143,144]
hPASk	kinase	*Homo sapiens*	NMR	1LL8	[160]
ARNT	translocator	*Homo sapiens*	NMR	1X0O	[161]
HIF2A	hypoxia inducible factor	*Homo sapiens*	X-ray	1P97	[162]
PER	clock protein	*Drosophila melanogaster*	X-ray	1WA9	[163]
HERG	K^+ channel	*Homo sapiens*	X-ray	1BYW	[15]
NCoA-1	coactivator	*Homo sapiens*	X-ray	1OJ5	[164]
Bphp	photoreceptor	*Deinococcus radiodurans*	X-ray	1ZTU	[165]

Table 7.1: **Details on the currently known structures of PAS domains.** For PYP, FixL and DOS many structures are available. Only the dark state structure of PYP and the unliganded structures of FixL and DOS are listed. These structures are also used in figure 7.1.

co-factors in the PAS core. All co-factors are located on the concave side of the β-sheet. The helices at the right-hand-side of the β-sheet accomodate the more bulky cofactors heme (FixL, DOS) and flavin (LOV). Situated more in the center of the protein are the smaller compounds: homoserine lactone (TraR), para-coumaric acid (PYP) and citrate (CitA). One PAS domain structure binds a small peptide (NCoA-1), located on top of the central α-helices. This peptide is an interaction motif, and represents the interaction of a PAS domain with another protein. The location of the cofactors in the PAS fold indicates that signal perception occurs in a conserved location.

To address the question whether signal communication also occurs in a conserved location, a more detailed description of the PAS domain function is required. PAS domains can exhibit two functions: (i) acting as a sensor domain for signal reception and (ii) mediating protein-protein interactions. In bacteria, the former is prevalent; many one component and two-component signal transduction systems contain a PAS domain as the sensory input domain [2]. In eukaryotes, PAS domains function as sensors and as dimerization domains, often N-terminal to basic helix-loop-helix (bHLH) motifs.

As a sensor domain, the PAS domain can detect a variety of signals, such as light or small molecules, often enabled by the incorporation of a co-factor. The PAS domain of FixL contains a heme, enabling the sensing of O_2, CO, CN and NO. FixL is part of a two-component system in a bacterial symbiont of rice (*Sinorhizobium meliloti* and *Bradyrhizobium japonicum*), that activates the expression of nitrogen fixation genes in low oxygen concentrations. The release of oxygen induces a change in planarity of the heme group, followed by a conformational change in the protein [166, 167]. Another bacterial oxygen sensor, DOS, is the sensory input domain to a phosphodiesterase (PDE) that linearizes cyclic di-GMP (c-di-GMP) in the absence of oxygen, inhibiting the production of cellulose. Here, the absence of oxygen induces a methionine side chain to bind to the heme group, rigidifying a loop in the protein [156]. Although both oxygen sensors utilize a heme group embedded within a PAS domain, their mechanisms of signal transduction differ, tuned by side chain interactions with the heme group. The different mechanisms may reflect the different types of adaptation: inhibition of a protein is a temperorary adaptation, while the activation of genes is chronic [12].

The inclusion of a light sensitive cofactor in a PAS domain enables photoreception. The structural protoype of the family, PYP, contains the chromophore p-coumaric acid, covalently linked to the polypeptide chain. Functionally, the protein is involved in the phototactic response of the bacterium *Halorhodospira halophila* to harmful blue light. Another PYP, from *Rhodospirillum centenum*, is the sensory domain of the histidine kinase Ppr, involved in a photoprotection response [158]. Upon the absorption of a blue light photon, the protein undergoes a series of structural and conformational rearrangements, that lead to a partially unfolded signaling state. Characteristic of this state is that it is partially unfolded and that the chromophore is expelled from the protein interior, (see Chapter 3). Within seconds the protein returns to its resting state (see Chapter 4).

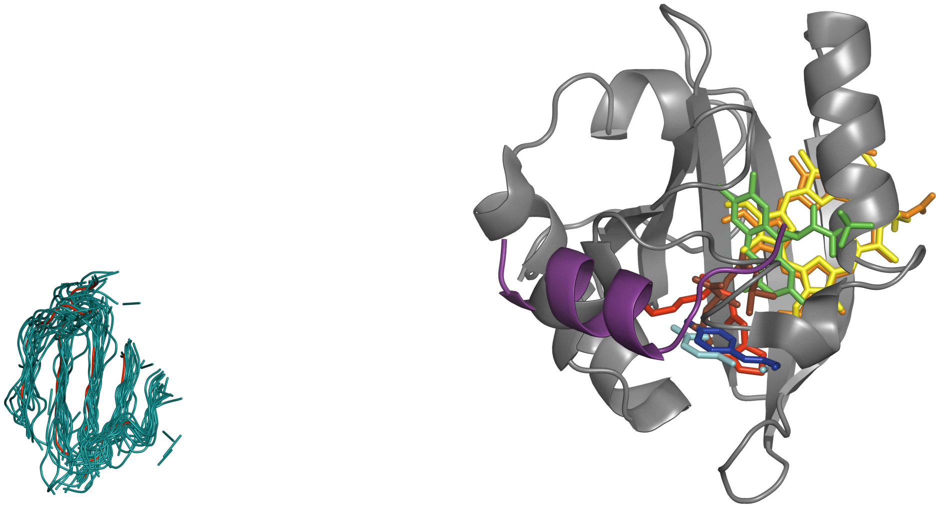
Higher organisms also employ PAS domains for light-sensing. Phototropins are light-activated kinases involved in plant responses to blue light. As N-terminal sensory input domains these proteins contain two PAS domains, called LOV1 and LOV2, linked to a C-terminal serine/threonine kinase domain. LOV is the abbreviation of Light-Oxygen-Voltage, a subclass within the PAS family. The LOV domain binds flavin as a cofactor for light detection. Upon light absorption, the flavin forms a covalent bond with the protein, but no further conformational changes have been observed in the PAS core [159, 168]. A combination of mutagenesis experiments and NMR spectroscopy has demonstrated that the unfolding of the Jα-helix, C-terminal to the LOV domain, is the critical event in the regulation of kinase activation [169].

PAS domains can detect some small molecules by incorporating them into their three dimensional fold. CitA is a bacterial citrate sensor from *Klebsiella pneumonia* located at the periplasmic side of the membrane and the first PAS domain implicated in signal transduction across the membrane. Citrate, when bound, interacts with the C-terminal β-strand of the PAS fold, a putative interaction site for the ligand [122]. TraR, a quorum sensor from *Agrobacterium tumefaciens*, uses a similar mechanism to perceive its signal. Bacteria use chemical signals to regulate gene expression at high population densities. Such a chemical signal induces the PAS core of TraR to fold, and furthermore stimulates the DNA binding domain of TraR to bind to DNA (see Chapter 6). The sensor for small organic molecules in PAS kinase has adopted a different way for sensing small compounds. The protein domain contains a flexible loop that facilitates the incorporation of these molecules, and this region serves as kinase binding site as well [160].

Besides the perception of signals PAS domains function also as dimerization domains. One example is the TraR system: the protein binds DNA as a dimer, and its PAS domains mediate the dimerization. Also basic helix-loop-helix (bHLH) motifs bind DNA as a dimer. Specificity is often regulated via a PAS domain or leucine zipper domain linked to the bHLH motif. Now, PAS domains function as mediator of protein-protein interactions. Class I bHLH-PAS proteins, including the aryl hydrocarbon receptor and the hypoxia inducible receptor, do not homo- nor heterodimerize with other class I transcription factors. For DNA-binding, heterodimerization with a class II bHLH-PAS, transcription factors that do homo- and heterodimerize, must occur. Aryl hydrocarbon nuclear receptor translocator (ARNT) is such a class II dimerization domain [170]. As an example, NPAS2, a gas responsive transcription factor found in mammals, has two PAS domains that both bind a heme group. When active, NPAS2 forms a heterodimer with a class II bHLH and binds DNA. Carbon monoxide molecules bound to the heme groups prevent the heterodimerization of NPAS2, thus blocks transcription [171]. Furthermore, swapping two PAS domains in a class I and class II bHLH protein may lead to a DNA binding complex, but does not lead to transcription, indicating that the PAS domains regulate the orientation of the proteins relative to each other and the DNA [170]. An NMR study on the PAS-B domain in hypoxia-inducible factor (HIF) indicates that the central β-sheet is an interface for protein-protein interactions [162]. This finding is further substantiated by a structural investigation of the ARNT PAS domain, and the dimerization characteristics of ARNT-PAS and HIF-PAS. Spin-labeling experiments have shown that the PAS domains interface via their β-sheets in an anti-parallel way. This is a different mechanism than observed for TraR. The PAS domains of this protein form a dimer mediated by helices N-terminal to the PAS domain. Also, the β-sheets in both monomers are oriented differently with respect to each other. As the PAS domain in TraR is also a sensory input domain, its function differs from the PAS domains mediating interactions in bHLH transcription factors.

The dimerization function of PAS domains extends to the formation of heterodimers. A human PAS domain was discovered in the sequence of a potassium channel called HERG (human ether-a-go-go related gene) [15] where it controls the current of K^+ through this channel via intra-protein interactions [16]. Mutations in HERG can cause a prolongation of cardiac repolarization, the clinical hallmark of the long QT syndrome, an inherited disorder with a propensity for syncope and arrhythmic sudden death. Alterations in the PAS domain can lead to defective trafficking of the protein to the cell membrane [17] or acceleration of channel deactivation. The PAS domain of HERG lacks a co-factor. Mutations in this PAS fold, causing the long QT syndrome, are located in the region where other PAS domains bind their cofactor [18, 19]. PAS domain interactions are also involved in the regulation of circadian rhythms. In *Drosophila melanogaster* the blue-light sensing cryptochromes interact with Per and its interaction partners Timeless and Doubletime to synchronize the internal clock with the environmental day-night cycle. PER contains two PAS domains, PAS-A and PAS-B, that bind Timeless and Doubletime. The structure of Per shows a non-crystallographic dimer with PAS-A accomodating a tryptophan residue of its dimerizing partner, whereas the β-sheet of PAS-B is implicated in mediating interactions with other proteins [163]. Yet another example of PAS domains interacting with various proteins comes from a transactivation process involved in an immune response. STAT is a transactivator and acts together with the interleukin transcription factor. When activated, STAT proteins recruit co-activators, including NCoA. The interaction between the transactivator STAT and co-activator NCoA is very specific and ocurs via an interaction motif. The structure of NCoA-1 in complex with the peptide representing this interaction motif is an example of interactions between a PAS domain and a different protein domain [164].

This overview ends with BphP, a red light receptor from the bacterium *Deinococcus radiodurans*. As a phytochrome, the protein contains several sensory input domains, including a PAS and a GAF domain and a kinase output domain. The PAS family exhibits structural similarities



(a) (b)

Figure 7.1: **The PAS fold and co-factors.** (a) All thirteen known PAS structures were similarly oriented using the central β-sheet as a reference. This figure shows the β-strands in tube representation, with the Photoactive Yellow Protein highlighted in red. (b) The backbone of PYP is shown in ribbon representation. The co-factors are displayed as a stick model, after fitting the PAS domain to PYP, using the β-sheet as reference point. Orange: FixL, Yellow: DOS, green: LOV2, blue: PYP from *Halorhodospira halophila*, cyan: PYP from *Rhodospirillum centenum*, red: TraR, brown: CitA, purple: NCoA-1. Note that this is not a co-factor but a interaction motif from another protein.

to the GAF domain [†], that contains an anti-parallel six-stranded β-sheet separating a group of four α-helices from one short α-helix. Superposition of the GAF domain of YKG9 (found in *Saccharomyces cerevisiae*) on PYP shows that the location of a putative active site of YKG9 coincides with the chromophore binding pocket of PYP [172]. The GAF domain of Bphp binds the chromophore, linear tetrapyrrole, at a similar position. As the interface between the PAS and GAF domain a stretch of amino acids N-terminal to the PAS domain passes through a loop in the GAF domain (that is C-terminal to the PAS domain). This knot is at the convex side of the central β-sheet of the PAS domain, and close to the chromophore binding site. In providing an additional interface between the PAS and GAF domains, the knot restricts the conformational freedom of the receptor to enable effient use of the (light) energy obtained during signal perception [165].

The observation that signal perception occurs at a conserved position in the PAS fold still holds. Further signal transduction seems to occur in different ways, depending on the nature of the signal and interactions with surrounding proteins. Recurring in all PAS domains described here is that the central β-sheet is stable and that the α-helices at the concave side are flexible. This agrees with the findings in Chapter 2 that PAS domains exhibit similar flexibilities when including only a minimal PAS core. Taking into account all details of a PAS domain, including extended loops and/or shortened helices, results in a differentiation of PAS domain dynamics, as observed in Chapter 5 and ref. [124].

[†] The GAF family is a protein family with cyclic-GMP-regulated cyclic nucleotide phosphodiesterases, adenylyl cyclases and the bacterial transcription factor FhlA

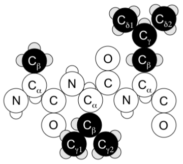
Appendix A

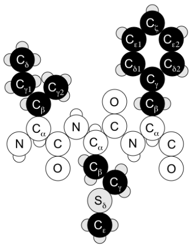
Protein structure

Amino acids are the primary building blocks of proteins. A typical amino acid contains an amino group, a carboxyl group, a hydrogen atom and a specific rest group, all bonded to a single carbon atom, C_α. Another name for the rest group is side chain. These side chains vary in size, shape, hydrogen bonding capacity, charges and chemical reactivity. The simplest amino acid is glycine, with just a hydrogen atom as rest-group, followed by alanine that has a methyl group as the side chain. All twenty amino acids together with their specific properties are listed in Table A.1.

Hydrophobic amino acids

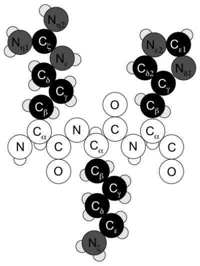
Alanine	Ala	A	Isoleucine	Ile	I
Valine	Val	V	Methionine	Met	M
Leucine	Leu	L	Phenylalanine	Phe	F

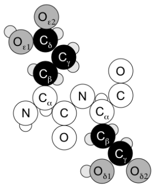




Charged amino acids

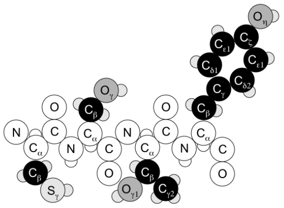
Arginine	Arg	R	Glutamate	Glu	E
Lysine	Lys	K	Aspartate	Asp	D
Histidine	His	H			

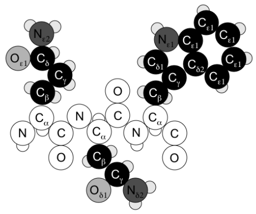




Hydrophilic amino acids

Cysteine	Cys	C	Glutamine	Gln	Q
Serine	Ser	S	Asparagine	Asn	N
Threonine	Thr	T	Tryptophan	Trp	W
Tyrosine	Tyr	Y			





Special amino acids

Glycine	Gly	G
Proline	Pro	P

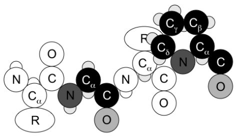


Table A.1: **The twenty amino acids commonly occurring in proteins.** Based on chemical properties, the amino acids can be divided into four groups: hydrophobic amino acids, hydrophilic amino acids, charged amino acids, and special amino acids - with properties that affect the orientation of the protein backbone. The R groups indicate side chains.

Linked by peptide bonds, amino acids form polypeptide chains. The side chains of amino acids in polypeptide chains are also called residues. Since the polypeptide chain has different ends, it has direction: the amino-terminal end is the beginning of a polypeptide chain, and is known as the N-terminus. The carboxyl end is called the C-terminus. In a polypeptide chain, the peptide bonds form the main chain, or backbone. Cysteine residues may form additional covalent linkages with another cysteine, known as disulfide bridges. Polypeptide chains spontaneously fold into three dimensional structures, governed by suitable peptide bond dihedral angles, hydrogen bond formation and hydrophobic interactions. These properties are closely related to the amino acid sequence, since the separate amino acids each have their unique properties.

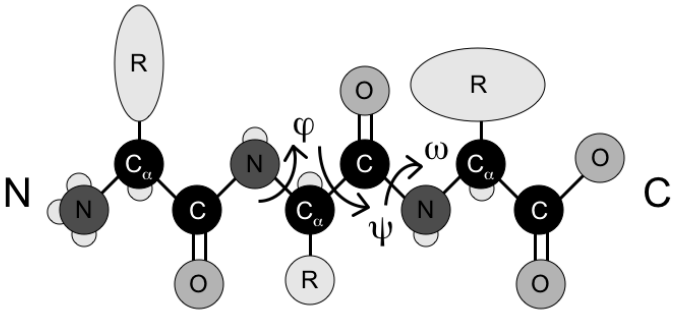


Figure A.1: **Primary structure of a protein.** The balls and sticks represent the protein backbone, including the N- and C-terminal ends. Residue side chains are displayed schematically. The arrows indicate the backbone torsion angles ϕ, ψ and ω.

Proteins are polypeptide chains that have a biological function. Their unique amino acid sequence is specified by a gene. Knowledge of the sequence of a protein is essential in understanding its mechanism of action. Furthermore, the protein sequence determines it three dimensional fold, governed by the interactions between the constituent atoms. Protein function relies on the binding of ligands and/or substrates and the transmission of structural changes. Proteins are unique in their capability to recognize and interact with a large diversity of molecules ranging from small compounds to other proteins and DNA. In order to perform their function, proteins visit different shapes specifically linked to their function. In conclusion, the amino acid sequence, the three dimensional structure and the function of a protein are intimately linked.

At first glance, a ball-and-stick model of a protein looks like a mass of atoms, without any obvious order. Simplifying the representation by removing the amino acid side chains shows the path of the polypeptide chain through space, revealing the secondary structure of a protein, see figure A.2. The secondary structure is organized in a small number of basic recurrent elements, such as helices and strands, connected by loops. When looking at the backbone of many proteins, more simplifying features emerge: helices and sheets form recurrent motifs and

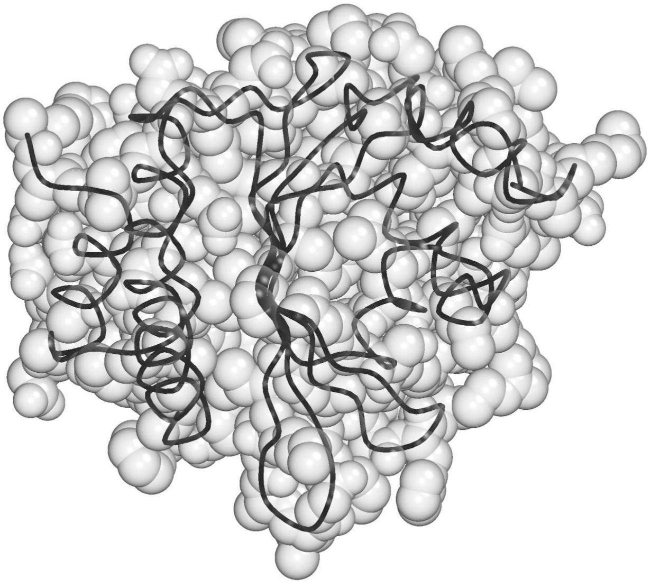


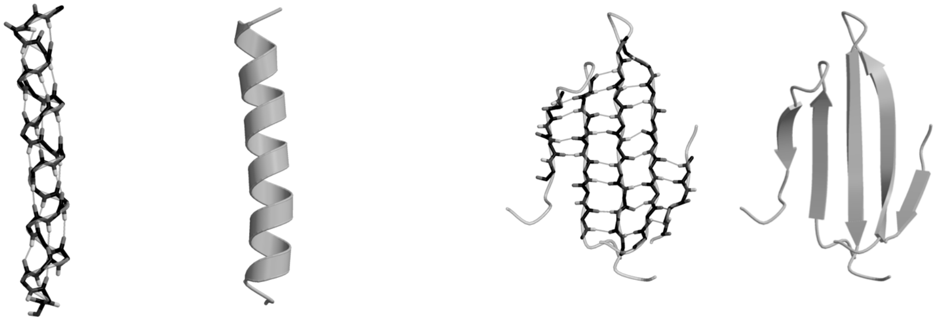
Figure A.2: **Atomic structure of a protein.** All atoms in this typical protein are displayed as spheres. The trace represents the protein backbone. Coordinates are taken from TraR, PDB-code 1L3L.

even structural domains. The underlying cause for the packing of a protein becomes obvious when the side chains and their mutual interactions are added back to the visualization model. Summarizing, a good way to analyze protein structure is by viewing it as a hierarchy of substructures. The commonly adopted view on protein structure [173] identifies four levels, with the amino acid sequence as the first level or primary structure. Hydrogen bonds between a carbonyl oxygen and an NH group in the backbone of the polypeptide chain cause the formation of ordered regions in a protein. Such regions occur as either an α-helix , figure A.3(a), or a β-sheet, figure A.3(b), and make up the secondary structure of a protein, together with loops and turns.

The C=O group of a residue i in an α-helix hydrogen bonds to the NH group of residue $i+4$. All C=O groups in the helix are arranged parallel to the long axis of the helix, while the side chains point away from this axis, resulting in an helical dipole. The first and the last residue in an α-helix are different from those in between, since part of the intrahelical hydrogen bonds cannot form. These residues require either interactions with solvent or other groups (*i.e.* side chains) in the protein. Another characteristic of α-helices is that they are often amphipathic: hydrophobic side chains are at one side, while hydrophilic residues are at the other side. A single extended stretch of amino acids, a β-strand, may interact with a similarly extended chain through the formation of hydrogen bonds between their backbone atoms. The two strands may then interact with additional chains, leading to the formation of a β-sheet. Two stable arrangements can occur: the parallel β-sheet and the anti-parallel β-sheet.

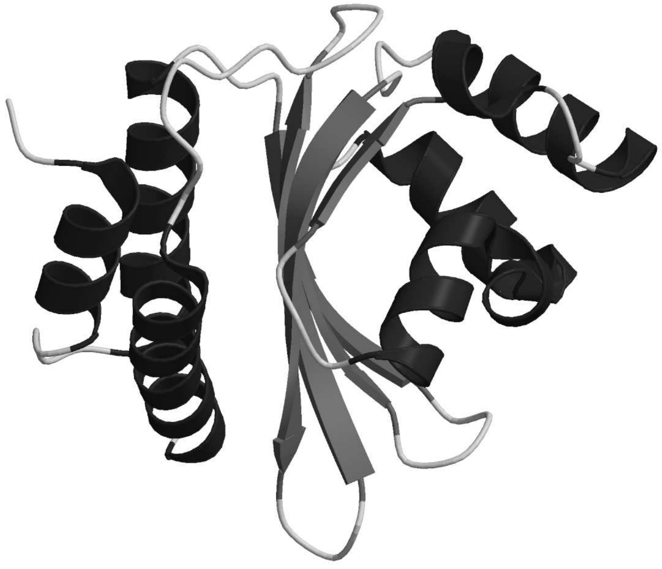
A polypeptide chain can form several of these structural elements and connect them via surface exposed loop regions. The orientation of the protein backbone, defined by the ϕ and

ψ dihedral angles, see figure A.1, specifies the secondary structure. Side chains restrict the rotation around the backbone bonds. All residues in an α-helix or a β-sheet adopt a specific conformation, with characteristic dihedrals.



(a) secondary: α-helix

(b) secondary: β-sheet



(c) tertiary

Figure A.3: **Schematic overview of protein structure.** Secondary structure elements, (a) α-helix and (b) β-sheet, are displayed in ball and stick representation and in green ribbon representation. In the first, backbone hydrogen bonds are highlighted as transparent bonds. (c) Tertiary structure - a fold comprising a central β-sheet in green and α-helices in red. Loops are displayed in grey. Coordinates are taken from PDB-code 1L3L.

Secondary structure elements often occur as a group, and are recognized as such as motifs and/or supersecondary structures; there are many possibilities to form β-sheets, but only a few of those occur regularly (such as the β-hairpin motif). The packing of a single polypeptide chain is known as the tertiary structure. The main driving force for the formation of the tertiary structure is the burial of hydrophobic side chains (minimizing their contact with water). This means that a folded protein is generally compact, and that buried hydrogen-bonding groups are paired. Such conditions favour the formation of helices and sheets, since the hydrogen bonding groups in the protein backbone are paired in these conformations, and these structures allow dense packing of hydrophobic side chains. Consequently, the large majority of buried charged groups and hydrogen bonding groups must be paired, or have a catalytic function. Loops connecting the ordered secondary structure elements are usually at the protein surface, exposed to solvent (aqueous solution). In general, globular proteins have a hydrophobic core with the charged groups at their surface.

Many proteins are only functional when comprising several subunits, not necessarily covalently linked. This aspect of spatial structure is known as the quaternary structure and includes the subunit arrangement and their contacts and interactions, see for example figure A.4. Usually the subunit interfaces in that region are as densely packed as any protein interior, and charged groups and hydrogen bonding groups are paired [8, 173].

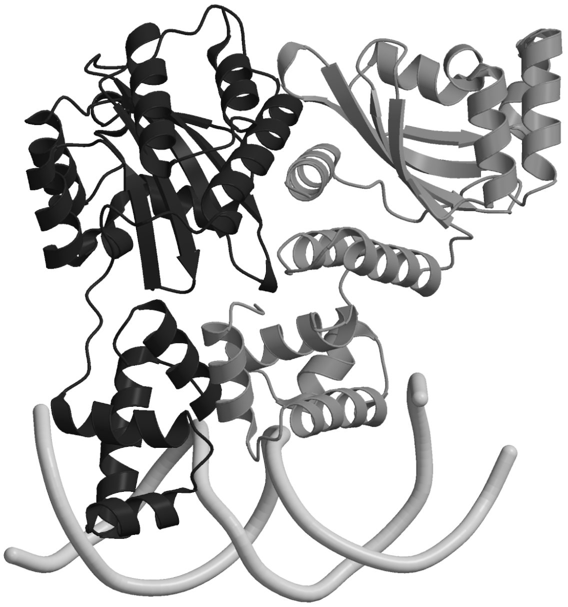


Figure A.4: **Protein quaternary structure.** Ribbon representation of a dimeric protein (green and red) in complex with DNA (blue and grey). Coordinates are taken from PDB-code 1L3L

Structure determination

Many experimental techniques underly our current knowledge of protein structure. Obtaining the primary protein structure is achieved through biochemical methods, e.g. through direct determination of amino acid sequence or, faster, the nucleotide sequence of the corresponding gene. Electron microscopy facilitates the determination of quaternary protein structure such in as ribosomes and other protein complexes, albeit at low resolution, lacking atomic detail. The determination of secondary and tertiary structure requires detailed information on the location of atoms within a protein. So far, the main methods to obtain three-dimensional structures at atomic resolution are X-ray crystallography and NMR spectroscopy. All protein structures can be deposited in the Protein DataBase (PDB, [119]).

X-ray crystallography

An essential prerequisite for protein structure determination using X-ray crystallography is a protein crystal that strongly diffracts X-rays. Within a crystal, repeating unit cells form a regular lattice that acts as a three dimensional diffraction grating to scatter X-rays. Such a unit cell may contain one or more protein molecules. X-ray radiation directed at a crystal interacts with the electrons in the crystal, causing them to oscillate. The oscillating electrons are a new source of X-rays, emitting in all directions: scattering. Diffraction spots occur where the scattered beams positively interfere with one another, ultimately resulting in a diffraction pattern. From the diffraction pattern, the distribution of the electron density in the unit cell of the crystal can be calculated using Fourier transformation. Each diffraction spot has an amplitude, the intensity of the spot, a wavelength, set by the X-ray source, and a phase. Information about the phase is lost during the experiment. Obtaining these phases, essential for the Fourier transformation, is the most difficult aspect of the determination of the structure, but there are several methods to solve this phase problem. The isomorphous replacement method uses the difference in scattering from crystals with and without heavy atoms included in the unit cell, such that the heavy atoms should not disturb the protein packing. Anomalous diffraction methods replace the sulphur in methionine by selenium, an anomalous scatterer with a distinctly different scattering, that allows determination of the phases. When a closely related structure is available, molecular replacement may result in the appropriate phase information. Once the phases are known, the distribution of electron density in the crystal unit cell can be calculated. Determination of the atomic structure requires interpretation of the electron density as a polypeptide chain with a particular amino acid sequence. Interpretation of the electron density map becomes more accurate with increasing resolution. Electron density maps at low resolution (4-6 Å) reveal the overall shape of the molecule, whereas they allow tracing of the peptide backbone at 3.5 Å. Fitting the amino acid sequence to the electron density is possible at a resolution of 3 Å or higher, whereas atomic resolution is possible below 1.5 Å. Building a structural model that fits the electron density is an iterative refinement process that aims to minimize the differences between the amplitudes calculated for the atomic model and the actual data, and simultaneously optimizes the model for the atomic packing in the unit cell.

Despite major breakthroughs in solving the phase problem, determining protein structures with X-ray crystallography suffers from a major drawback: the need to obtain well-ordered protein crystals. Since proteins are large and globular molecules, packing them into a lattice results in the formation of large holes or channels between the molecules. Such channels can occupy more than half of the volume of the crystal and are usually filled with disordered solvent molecules. The success of packing proteins into a regular lattice is heavily dependent on various external parameters (pH, temperature, concentration of the protein, co-crystallizers, ions, et cetera). Nevertheless, X-ray crystallography has been the main source for protein structures.

NMR spectroscopy

Applied to moderately sized proteins NMR spectroscopy enables the determination of the structures of proteins in solution. Certain atomic nuclei, e.g. ^{1}H, ^{13}C, ^{15}N and ^{31}P, have a magnetic moment. The ^{1}H isotope is naturally abundant in proteins, but growing micro-organisms on media enriched with ^{13}C and/or ^{15}N allows incorporation of these isotopes in proteins too. Placing molecules containing these isotopes in a magnetic field causes their spin to align along the field. This equilibrium environment can be excited with radio frequency pulses. The molecules return to the equilibrium state via emission of radio frequency radiation, that can be measured. The frequencies of the emitted waves depend on the chemical environment of the nucleus. When obtained relative to a reference signal these frequencies are called chemical shifts. Using a variety of pulses allows probing several properties of individual spins. Nuclei can interact through chemical bonds (spin-spin coupling) and through space (Nuclear Overhauser Effect, NOE). In principle, each nucleus with a magnetic moment gives a unique signal (except for chemically equivalent atoms), but the huge amount of atoms in a protein molecule prohibits their straightforward determination. Using advanced two-dimensional techniques and a variety of pulse sequences allows the assignment of individual protons, carbon- and nitrogen atoms. Using the NOE, protons within close distance can be determined. The distances between atom pairs are then used as restraints in energy calculations to obtain an ensemble of structures that fit these restraints. Within the ensemble, the structures can vary, depending on the amount and consistency of the information.

Structure prediction

Genome projects have provided the complete sequence of all the genes in many organisms, including *Homo sapiens*. Assigning a function to all the gene products, the proteins, often requires knowledge of protein structure. Interactions at the atomic level determine the fold of a protein; hydrogen bonds between backbone atoms result in the formation of secondary structure elements and side chain interactions govern the packing of these structure elements, leading to the tertiary structure, the fold, of a protein. Nevertheless, in principle, the three dimensional shape of a protein is already defined in its amino acid sequence. In 1968 Levinthal argued that since a protein with a specific amino acid sequence has an astronomical number of conformations to choose from, an unbiased search through all of them would take too long [174, 175]. This is certainly true when trying to predict the fold of a protein from its amino acid sequence. Fortunately, including additional information about the protein greatly improves our ability to predict its structure.

Many functional assignments are based on similarity with proteins of known function. Based on function, proteins are grouped into protein families, that exhibit significant similarity at different levels of protein structure. Such similarity is known as homology when the proteins under comparison have evolved from the same ancestral gene. Homologous proteins have significant similarities in their amino acid sequence and tertiary fold. To find the function of a protein, comparing its sequence against a database of protein sequences of known function is a useful tool, as provided by BLAST [176] and Pfam [177]. When two proteins have a significant number of identical amino acids at corresponding positions in their polypeptide chain, with randomly generated sequences as a reference, they are homologous. Homology occurs also on higher levels of protein structure: Sharing a similar fold might indicate that the protein performs a similar function. Identifying key functional residues then improves the recognition of subsequent homologous proteins. After identification of a protein sequence, prediction of the three dimensional structure of this protein can use the homology to a known fold. The DALI server [47] is a useful tool to perform structural alignments.

Comparing tertiary folds of homologous proteins shows that similarity is highest in the hydrophobic core, built up by secondary structure elements, known as the scaffold. The loop regions that connect the strands and helices can vary in length, as well as in conformation. After establishing the structural scaffold, the remaining problems comprise predicting the conformation of loop regions and determination of energetically favorable side chain orientations. Prediction of the structure of loop regions depends on the number of amino acids in the region, and whether they connect two α-helices, two β-strands or a combination of a strand and a helix.

Appendix B

Protein Dynamics

Proteins have a significant degree of flexibility; for example, some enzymes and receptors change conformation upon the binding of a ligand or substrate. These conformational changes are as small or as large as relevant for the function of the protein. Proteins express several different modes of motion and flexibility. In solution, globular proteins rotate, a motion known as molecular tumbling. Within a protein, regions may expand and allow solvent molecules to penetrate into the core. When the region resumes its previous size, these solvent molecules are forced out again. This is called protein breathing. Contained within a protein, motion of atoms and side chains occurs, such as side chain rotation. In globular proteins, residues at the protein surface have more degrees of freedom than residues in the protein core. Specific motions might also involve protein domains moving relative to each other, known as segmental flexibility [173].

These motions originate from atoms in a protein system interacting with each other, exchanging potential and kinetic energy. At room temperature atoms in an ideal gas have kinetic energy, proportional to their squared velocities:

$$< \frac{1}{2}mv^2 >= \frac{1}{2}k_BT \tag{B.1}$$

with m and v the masses and velocities of the atoms, k_B the Boltzmann constant and T the temperature.

In a classical description of a protein system, the potential energy is the sum of simple, empirically derived functions that describe the interactions between atoms. There are two classes of interactions within a protein system: bonded and non-bonded interactions. The first type includes all interactions between atoms that are connected via one or more covalent bonds (U_{bond}, U_{angle} and $U_{dihedral}$), whereas the second type describes interactions of atoms through space: the electrostatic and van der Waals interactions.

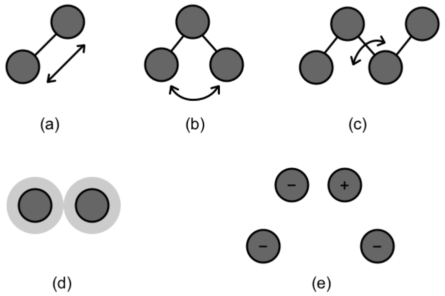
Van der Waals interactions comprise two processes: the penetration of an electron cloud by the electron cloud of another atom (scaling as r^{-12} and a repulsive interaction) and London dispersion forces, *i.e.* fluctuations in the electron density cause the formation of temporary dipoles, inducing the formation of dipoles in nearby atoms (scaling as r^{-6} and an attractive interaction). These processes combine into the following potential, known as the Lennard-Jones potential:

$$U_{LJ} = \frac{C_{ij}^A}{r^{12}} - \frac{C_{ij}^B}{r^6} \tag{B.2}$$

where r is the distance between two atoms. The minimum of this function corresponds to the van der Waals contact distance [173] and differs for different atom types, as indicated by C_A and C_B.

Electrostatic interactions are composed of multiple contributions: ion-ion $1/\epsilon r$, ion-dipole $1/\epsilon r^4$ and dipole-dipole $1/\epsilon r^6$, with r the distance between two charged groups and ϵ the dielectric constant. These interactions have a long range. The dielectric constant is very relevant for the strength of the interactions, it is 80 in water and in between 2 and 4 in the interior of a protein. Charged groups in the protein can facilitate higher values of ϵ within a protein, when they function as a lens to (de)focus charges and/or align water molecules [173]. Figure B.1 lists the interaction types that occur in a protein. The total potential energy of a protein system can be expressed as the sum of all these interactions:

$$U = U_{bond} + U_{angle} + U_{dihedral} + U_{LJ} + U_{Coulomb} \tag{B.3}$$



$$U_{bond} = \frac{1}{2}k_b^{ij}(r_{ij} - r_{ij}^0)^2$$
$$U_{angle} = \frac{1}{2}k_a^{ijk}(\theta_{ijk} - \theta_{ijk}^0)^2$$
$$U_{dihedral} = \frac{1}{2}k_d^{ijkl}(1 + cos(n\phi_{ijkl} - \phi_{ijkl}^0))$$
$$U_{LJ} = (\frac{C_{ij}^A}{r_{ij}})^{12} - (\frac{C_{ij}^B}{r_{ij}})^6$$
$$U_{coulomb} = \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}$$

Figure B.1: **Interactions between atoms in a protein system.** A harmonic potential describes the bond (a) and angle (b) interactions, and a periodic function describes the dihedral interaction (c), with k_i the respective force constants, r_{ij} the bond length between atoms i and j, θ_{ijk} the bond angle between atoms i, j and k and ϕ_{ijkl} the dihedral between atoms i, j, k and l. (d) The Lennard-Jones potential, U_{LJ}, describes the Van der Waals interaction between two atoms, with C_{ij}^A the repulsion coefficient, C_{ij}^B the attraction coefficient and r_{ij} the distance between atoms i and j. (e) Electrostatic interactions are described by the Coulomb equation, with q_i the (partial) charge on atom i, r_{ij} the distance between atoms i and j, and ϵ_0 the dielectric constant. These functions are taken from the GROMACS manual [178].

Monitoring the dynamics of proteins at atomic resolution requires advanced techniques, such as NMR spectroscopy and crystallography (See appendix A for a description of these techniques). Using molecular simulation methods in combination with sufficient computing power, computational protein models now complement experiment and even make predictions that are at high, atomic resolution [178].

Molecular Dynamics: the algorithm

Molecular dynamics (MD) is a popular method in the field of biomolecular simulation since it provides the temporal evolution of motions in large (atomistic) detail. The method resembles an experiment in many ways: it starts with setting up a system for simulation, followed by the actual measurements.

Setting up a simulation consists of selecting an appropiate starting configuration and equilibration of this configuration. This is also called initialization and starts with assigning positions and velocities to all particles. Initial velocities are taken from a distribution that reflects a given temperature. When simulating protein systems, structures obtained from crystallography or NMR spectroscopy usually provide the initial positions of the atoms. Such starting structures can contain errors resulting in unfavorable interactions in a system, see appendix A for more details. Also, depending on the aim of the simulation, additional atoms are required (such as hydrogen atoms, water molecules and ions). A procedure to minimize the potential energy in the system reduces the time required for equilibration. Molecular simulations aim to provide information on the bulk properties of a system, but simulating each particle in an experimental setup is impossible. However, mimicking experiments by putting a feasible number of particles in a box (walls of frozen particles) suffers from finite size artefacts. A way to overcome this problem is to use periodic boundary conditions: the simulation box is a cell in an infinite lattice. In such a system, considering only the interactions within a boxlength circumvents the need to evaluate an infinite number of interactions.

A system is equilibrated when the properties of the system no longer change with time. To measure an observable quantity in an MD simulation, it must be expressed as a function of the positions and momenta of the particles in the system. As an example, to measure the temperature in the simulation, the kinetic energy in the system is computed, using equation B.1. In a simulation, this equation is an operational definition of the temperature. Since the kinetic energy of the system fluctuates, so does the instantaneous temperature:

$$T(t) = \sum_{i=1}^{N} \frac{m_i v_i^2(t)}{k_B N_f} \tag{B.4}$$

with $N_f = 3N - 6$, and N the number of particles. An accurate estimate of the temperature requires averaging over many fluctuations.

Once the system is equilibrated, the actual sampling can begin. This comprises the following steps: Calculation of the forces acting on each particle in the system and integration of the equations of motion.

For the calculation of the forces in the system the distances between particles need to be evaluated, see figure B.1 for an overview of interactions in a protein:

$$f(r) = -\frac{dU(r)}{dr} \tag{B.5}$$

Each particle feels a force that is exerted by all other particles. Since it is not possible to calculate the force acting on one particle as a result of the interactions with all other particles in the system, the force acting on one particle is approximated by pairwise additive interactions. This means that $N(N-1)/2$ pair distances must be evaluated, scaling as N^2. As a consequence, the calculation of the forces acting on each particle in an MD simulation is the most time consuming part. After calculating the forces acting on each particle, integration of the equations of motion is the next step. One of the simplest algorithms to do this is called the Verlet algorithm [179].

The derivation starts with a Taylor expansion of the particle coordinate (r) around time t:

$$r(t+\Delta t) = r(t) + v(t)\Delta t + \frac{f(t)}{2m}\Delta t^2 + \frac{\Delta t^3}{3!}r''' + O(\Delta t^4) \tag{B.6}$$

and

$$r(t-\Delta t) = r(t) - v(t)\Delta t + \frac{f(t)}{2m}\Delta t^2 - \frac{\Delta t^3}{3!}r''' + O(\Delta t^4) \tag{B.7}$$

where $O(\Delta t^4)$ represents the fourth term and higher in the Taylor expansion. Summing these equations results in:

$$r(t+\Delta t) + r(t-\Delta t) = 2r(t) + \frac{f(t)}{m}\Delta t^2 + O(\Delta t^4) \tag{B.8}$$

that can be rearranged as

$$r(t+\Delta t) = 2r(t) - r(t-\Delta t) + \frac{f(t)}{m}\Delta t^2 + O(\Delta t^4) \tag{B.9}$$

The estimation of the new position contains an error in the order of $O(\Delta t^4)$. To obtain the velocities:

$$r(t+\Delta t) - r(t-\Delta t) = 2v(t)\Delta t \tag{B.10}$$

$$v(t) = \frac{r(t+\Delta t) - r(t-\Delta t)}{2\Delta t} + O(\Delta t^2) \tag{B.11}$$

with an error in the order of $O\Delta t^2$, larger than the error in the calculation of the positions. Since the velocities are recalculated every timestep from the more accurate positions, the error does not accumulate.

An alternative to the Verlet algorithm is the Leapfrog algorithm [180]. This algorithm evaluates the velocities at half-integer time steps to use them for the computation of the new positions. It is called leapfrog because the positions and velocities leap over each other's backs.

$$r(t+\Delta t) = r(t) + \Delta t v(t+\Delta t/2) \tag{B.12}$$

Updating the velocities:

$$v(t+\Delta t/2) = v(t-\Delta t/2) + \Delta t\frac{f(t)}{m} \tag{B.13}$$

This scheme gives identical trajectories as the Verlet algorithm. Since the positions and velocities are updated at $t+\Delta t$ and $t+\Delta t/2$, the kinetic and potential energies are calculated at different moments.

Figure B.2 shows a schematic representation of a molecular dynamics simulation.

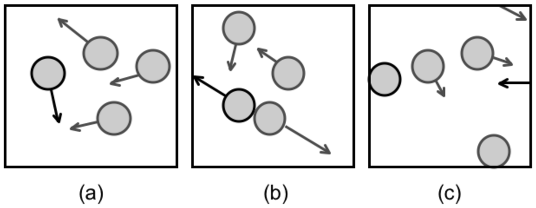


Figure B.2: **Schematic representation of a Molecular Dynamics simulation.** Frame (a) is the starting frame, where each particle has an initial position and velocity. Frame (b) and (c) show the positions and velocities after small time intervals, after which new forces are calculated. Frame (c) shows the effect of periodic boundary conditions: a particle that leaves the box, enters it simultaneously at the opposite side.

Speeding up simulations

Simulations at atomic detail of protein systems involve many particles and therefore long simulation times. Reducing the number of pairwise evaluations will reduce the CPU-time needed per timestep. When calculating interactions of a particle with all other particles, those located far from the particle of interest contribute little to the force acting on it. Truncating the interaction potential means that interactions of particles ranging beyond a specified distance are not considered. There are several possible ways of truncation: Simple truncation; Ignore all interaction beyond a cut-off radius r_c and shifted truncation: The potential is shifted such that it vanishes at r_c. The concept of the cut-off radius can be extended further, to speed up simulations even more. Using a basic cut-off radius requires the evaluation of all pairwise distances between the particles in the system to determine which interactions should be included in the calculation of the forces. Using a neighborlist reduces the number of these distance evaluations in the following way. Before the calculation of the interactions, a list is constructed of all particles within a radius r_n of particle i. During the calculation of the interactions, only the particles in the list have to be considered. Then, if the maximum displacement of the particles is less than the difference between the neighborlist radius r_n and the cut-off radius r_c, only the particles in the neighborlist are considered for the force calculation. If a particle moves farther than $r_n - r_c$ the neighborlist requires an update.

As discussed in the section on interactions between protein atoms, electrostatic interactions are long ranged. Using a cut-off on electrostatic interactions to speed up the simulation would exclude long range interactions, relevant for protein stability. Moreover, the electrostatic interactions may range beyond the simulation box, since it decays as $1/r$. For homogeneous systems, the Coulomb interaction can be modified by assuming a constant dielectric environment beyond the cut-off. For charged groups Coulomb interaction with reaction field corresponds to neutralization with a homogeneous background charge. The Ewald sum provides another solution. Assume that each particle i with charge q_i is surrounded by a diffuse charge distribution of the opposite sign, such that the total charge of the cloud cancels (screens) q_i. Now only the fraction of q_i that is not screened contributes to the effective electrostatic potential. This decays quite fast at large distances. A charge distribution that compensates for the screening charge distribution corrects for the addition of the screening charge cloud, to obtain point charges again. This distribution turns out to be continuous and periodic, enabling representation by a rapidly converging Fourier series. Then, instead of considering each point charge separately, distributing them on a mesh significantly speeds up the Fourier transformation of the Ewald sum. This method is also known as Particle Mesh Ewald [181, 182].

Another way of speeding up molecular dynamics simulations is the use of constraints on bonded interactions, usually bonds. For protein simulations, fluctuation of bond lengths adds little information, but requires very small time steps. A commonly used algorithm for constraining bonds is LINCS [178]. In general, the scheme works as follows. After an unconstrained update of the positions, the bonds are reset to the preferred values: first the projections of the new bonds on the old bonds are set to zero and then a correction is applied to the new bonds to account for rotation of the bonds. Such constraints fix the bond lengths, facilitating much larger time steps.

The level of detail in a force field depends on the issue of interest, and is a trade-off between simulation time and temporal/spatial resolution. A slower process requires more simulation time, but at the cost of detail, simulation time can be decreased. How many particles are required to adequately represent a macroscopic system? This is in the order of Avogadro's Number, 6.02310^{23} and far too large to handle even for the computer power that is currently available. Employing periodic boundary conditions allows sampling bulk phase with a small number of atoms (ranging from a few hundred to a few million).

Constant volume, temperature and pressure

The MD method as discussed in this appendix is constant in number of particles, volume and energy (constant-NVE). In contrast, real experiments are usually performed at constant temperature and constant pressure (constant-NpT) or at constant volume (constant-NVT). Keeping the volume constant is simply keeping the box size fixed. Fixing the temperature is more complicated. Many algorithms exist to keep the temperature fixed, and these can be divided into isokinetic and canonical schemes [82]. Isokinetic schemes use velocity-scaling to keep the average kinetic energy per particle constant (Berendsen thermostat [97]). Two possibilities exist for the second type of thermostat: the Andersen thermostat [183], where stochastic collisions with a heat bath of the desired temperature impose the temperature in the simulation system, and the Nosé-Hoover thermostat [130, 131], that uses extended equations of motion to set the temperature. For imposing constant pressure in a molecular dynamics simulation similar algorithms exist: the Berendsen barostat [97] uses scaling of the coordinates and box vectors to correct the pressure, whereas Parrinello-Rahman pressure coupling [128, 129] uses extended equations of motion to deform the box.

Parallel Tempering

Molecular Dynamics simulations, using a force field as described at the beginning of this chapter, are limited to relatively small time scales in the order of nanoseconds. Many processes that involve conformational changes in proteins are in the order of milliseconds and higher, caused by free energy barriers between metastable states. Studying protein dynamics with MD at constant temperature suffers from the restriction imposed by these barriers: the simulated protein cannot escape from its starting energy well. Increasing the temperature, while keeping the volume fixed, allows crossing of these energy barriers, expanding the conformational space sampled during the simulation. Cooling down the simulation results in the population of additional local minima, improving the sampling. This method is known as simulated annealing. Since the heating up and cooling down is done manually, no ensemble averages can be defined in such simulations.

The description of the simulated annealing method serves as a starting point to explain Parallel Tempering (PT), also known as Replica Exchange MD (REMD) [79–81]. Consider two MD simulations, replicas, running at a low temperature (i) and a high temperature (j), in a constrained volume. The high temperature replica crosses free energy barriers, while the low temperature replica samples the local minimum. Multiple exchanges of temperatures between these systems enable the crossing of free energy barriers and the sampling of local minima.

A typical parallel tempering simulation has M replicas, each in the NVT ensemble, at different temperatures (T_i). Since the replicas do not interact energetically, the partition function of this extended ensemble is:

$$Q = \prod_{i=1}^{M} \frac{1}{\Lambda_i^{3N} N!} \int d\mathbf{r_i^N} \exp(-\beta_\mathbf{i} \mathbf{U}(\mathbf{r_i^N})) \tag{B.14}$$

with $\mathbf{r_i^N}$ the positions of the N particles in replica i and Λ^{3N}, representing the temperature dependent part of replica i ($\Lambda^{3N} = \prod_{i=1}^{N} (2\pi m_j k_B T_i)^{3/2}$). Attempts at exchanging temperatures between two replicas are governed by a Monte Carlo scheme. Such a scheme obeys detailed balance; the probability of the forward swap ($i \rightarrow j$) must be equal to the reversed swap ($j \rightarrow i$):

$$W(old)acc(old \rightarrow new) = W(new)acc(new \rightarrow old) \tag{B.15}$$

$W(old)$ and $W(new)$ are given by the Boltzmann distribution and result in the following acceptance rule:

$$\frac{acc(o \rightarrow n)}{acc(n \rightarrow o)} = \frac{W(n)}{W(o)} = \frac{i \rightarrow j}{j \rightarrow i} = \frac{\exp(-\beta_i U(\mathbf{r_j}) - \beta_\mathbf{j} \mathbf{U}(\mathbf{r_i}))}{\exp(-\beta_i U(\mathbf{r_i}) - \beta_\mathbf{j} \mathbf{U}(\mathbf{r_j}))} \tag{B.16}$$

The Metropolis rule then gives:

$$P_{acc}(ij) = min\left(1, e^{\Delta\beta_{ij}\Delta U_{ij}}\right) \tag{B.17}$$

with $\Delta\beta_{ij}$ the difference of the inverse of the swapping temperatures and ΔU_{ij} the energy difference of the two systems. To enable exchange between the systems, the potential energies must have sufficient overlap. This is tuned via the size of the energy gap. Usually, this means that more than two replicas are required to achieve the crossing of free energy barriers, while probing the local minima.

Bibliography

[1] J.A. Hoch and T.J. Silhavy, editors. *Two-Component Signal Transduction*. ASM Press, Washington DC, 1995.

[2] L.E. Ulrich, E.V. Koonin, and I.B. Zhulin. One-component systems dominate signal transduction in prokaryotes. *Trends Microbiol.*, 13:52–56, 2005.

[3] J. Parkinson. *Genetic approaches for signaling pathways and proteins*, pages 9–24. ASM Press, Washington DC, 1995.

[4] J.B. Stock, M.G. Surette, M. Levit, and P. Park. *Two-component signal transduction systems: Structure-function relationships and mechanisms of catalysis*, pages 9–24. ASM Press, Washington DC, 1995.

[5] A. Khorchid and M. Ikura. Bacterial histidine kinase as signal sensor and transducer. *Int. J. Biochem. Cell Biol.*, 38:307–312, 2006.

[6] J.A. Hoch. Two-component and phosphorelay signal transduction. *Curr. Opin. Microbiol.*, 3:165–170, 2000.

[7] L. Stryer. *Biochemistry*. W.H. Freeman and Company, New York, fourth edition, 1995.

[8] C. Branden and J. Tooze. *Introduction to protein structure*. Garland Publishing Inc., New York, second edition, 1999.

[9] B. Alberts, D. Bray, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Essential cell biology*. Garland Publishing, Inc., New York, 1998.

[10] B. Hao, C. Isaza, J. Arndt, M. Soltis, and M.K. Chan. Structure-based mechanism of O_2 sensing and ligand discrimination by the fixl heme domain of *Bradyrhizobium japonicum*. *Biochemistry*, 41:12952–12958, 2002.

[11] L.E. Cybulski, D. Albanesi, M.C. Mansilla, S. Altabe, P.S. Aguilar, and D. de Mendoza. Mechanism of membrane fluidity optimization: Isothermal control of the *Bacillus subtilis* acyl-lipid desaturase. *Mol. Microbiol.*, 45:1379, 2002.

[12] M.A. Gilles-Gonzalez and G. Gonzalez. Signal transduction by heme-containing PAS-domain proteins. *J. Appl. Physiol.*, 96:774–783, 2004.

[13] J.R. Tuckerman, G. Gonzalez, E.M. Dioum, and M.A. Gilles-Gonzalez. Ligand and oxidation-state specific regulation of the heme-based oxygen sensor FixL from *Sinorhizobium meliloti*. *Biochemistry*, 41:6170–6177, 2002.

[14] C.P. Ponting and L. Aravind. PAS: A multifunctional family comes to light. *Curr. Biol.*, 7:R674–R677, 1997.

[15] J.H.M. Cabral, A. Lee, S.L. Cohen, B.T. Chait, M. Li, and R. Mackinnon. Crystal structure and functional analysis of the HERG potassium channel N terminus: A eukaryotic PAS domain. *Cell*, 95:649–655, 1998.

[16] D. Gómez-Varela, P. de la Peña, J. García, T. Giráldez, and F. Barros. Influence of amino-terminal structures on kinetic transitions between several closed and open states in human *erg* K^+ channels. *J. Membrane Biol.*, 187:117–133, 2002.

[17] A. Paulussen, A. Raes, G. Mattijs, D.J. Snyders, N. Cohen, and J. Aerssen. A novel mutation in the PAS domain of the human potassium channel HERG results in the long QT syndrome by trafficking deficiency. *J. Mol. Biol.*, 277:48610–48616, 2002.

[18] J. Chen, A. Zou, I. Splawski, M.T. Keating, and M.C. Sanguinetti. Long QT syndrome associated mutations in the Per-Arnt-Sim (PAS) domain of HERG potassium channels accelerate channel deactivation. *J. Mol. Biol.*, 274:10113–10118, 1999.

[19] L. Shushi, B. Kerem, M. Goldmit, A. Peretz, B. Attali, A. Medina, Towbin. J.A., J. Kurokawa, R.S. Kass, and J. Benhorin. Clinical, genetic and electrophysiologic characteristics of a new PAS-domain HERG mutation M124R causing long QT syndrome. *Ann. Noninvasive Electrocardiol.*, 10:334–341, 2005.

[20] J.L. Pellequer, K.A. Wager-Smith, S.A. Kay, and E.D. Getzoff. Photoactive Yellow Protein: A structural prototype for the three-dimensional fold of the PAS domain superfamily. *Proc. Natl. Ac. Sci. USA*, 95:5884–5890, 1998.

[21] B.L. Taylor and I.B. Zhulin. PAS domains: Internal sensors of oxygen, redox potential, and light. *Microbiol. Mol. Biol. Rev.*, 63:479–506, 1999.

[22] I.B. Zhulin and B.L. Taylor. PAS domain S-boxes in Archaea, bacteria and sensors for oxygen and redox. *Trends In Biochemical Sciences*, 22:331–333, 1997.

[23] J.R. Nambu, J.O. Lewis, K.A. Wharton, and S.T. Crews. The Drosophila single-minded gene encodes a helix-loop-helix protein that acts as a master regulator of CNS midline development. *Cell*, 67:1157–1167, 1991.

[24] V. Anantharaman, E.V. Koonin, and L. Aravind. Regulatory potential, phyletic distribution and evolution of ancient, intracellular small-molecule-binding domains. *J. Mol. Biol.*, 307:1271–1292, 2001.

[25] W.W. Sprenger, W.D. Hoff, J.P. Armitage, and K.J. Hellingwerf. The eubacterium *Ectothiorhodospira halophila* is negatively phototactic, with a wavelength dependence that fits the absorption-spectrum of the Photoactive Yellow Protein. *J. Bacteriol.*, 175:3096–3104, 1993.

[26] G.E.O. Borgstahl, D.R. Williams, and E.D. Getzoff. 1.4 ångstrom structure of Photoactive Yellow Protein, a cytosolic photoreceptor - Unusual fold, active-site, and chromophore. *Biochemistry*, 34:6278–6287, 1995.

[27] A.H. Xie, W.D. Hoff, A.R. Kroon, and K.J. Hellingwerf. Glu46 donates a proton to the 4-hydroxycinnamate anion chromophore during the photocycle of Photoactive Yellow Protein. *Biochemistry*, 35:14671–14678, 1996.

[28] B. Perman, V. Srajer, Z. Ren, T.Y. Teng, C. Pradervand, T. Ursby, D. Bourgeois, F. Schotte, M. Wulff, R. Kort, K. Hellingwerf, and K. Moffat. Energy transduction on the nanosecond time scale: Early structural events in a xanthopsin photocycle. *Science*, 279:1946–1950, 1998.

[29] R. Kort, H. Vonk, X. Xu, W.D. Hoff, W. Crielaard, and K.J. Hellingwerf. Evidence for trans-cis isomerization of the p-coumaric acid chromophore as the photochemical basis of the photocycle of Photoactive Yellow Protein. *FEBS Lett.*, 382:73–78, 1996.

[30] U.K. Genick, G.E.O. Borgstahl, K. Ng, Z. Ren, C. Pradervand, P. Burke, V. Srajer, T. Teng, W. Schildkamp, D.E. McRee, K. Moffat, and E.D. Getzoff. Structure of a protein photocycle intermediate by millisecond time-resolved crystallography. *Science*, 275:1471–1475, 1997.

[31] U.K. Genick, S.M. Soltis, P. Kuhn, I.L. Canestrelli, and E.D. Getzoff. Structure at 0.85 ångstrom resolution of an early protein photocycle intermediate. *Nature*, 392:206–209, 1998.

[32] D.M.F. van Aalten, W.D. Hoff, J.B.C. Findlay, W. Crielaard, and K.J. Hellingwerf. Concerted motions in the Photoactive Yellow Protein. *Protein Eng.*, 11:873–879, 1998.

[33] D.M.F. van Aalten, W. Crielaard, K.J. Hellingwerf, and L. Joshua-Tor. Conformational substates in different crystal forms of the Photoactive Yellow Protein - Correlation with theoretical and experimental flexibility. *Protein Sci.*, 9:64–72, 2000.

[34] D.M.F. van Aalten, A. Haker, J. Hendriks, K.J. Hellingwerf, L. Joshua-Tor, and W. Crielaard. Engineering photocycle dynamics - Crystal structures and kinetics of three Photoactive Yellow Protein hinge-bending mutants. *J. Biol. Chem.*, 277:6463–6468, 2002.

[35] M.A. van der Horst, I.H. van Stokkum, W. Crielaard, and K.J. Hellingwerf. The role of the N-terminal domain of Photoactive Yellow Protein in the transient partial unfolding during signalling state formation. *FEBS Lett.*, 497:26–30, 2001.

[36] W.M. Gong, B. Hao, S.S. Mansy, G. Gonzalez, M.A. Gilles-Gonzalez, and M.K. Chan. Structure of a biological oxygen sensor: A new mechanism for heme-driven signal transduction. *Proc. Natl. Ac. Sci. USA*, 95:15177–15182, 1998.

[37] S. Crosson and K. Moffat. Structure of a flavin-binding plant photoreceptor domain: Insights into light-mediated signal transduction. *Proc. Natl. Ac. Sci. USA*, 98:2995–3000, 2001.

[38] Z. Otwinowski and W. Minor. Processing of X-ray diffraction data collected in oscillation mode. *Macromol. Cryst. A*, 276:307–326, 1997.

[39] J. Navaza. Amore - An automated package for molecular replacement. *Acta Cryst. A*, 50:157–163, 1994.

[40] A.T. Brunger, P.D. Adams, G.M. Clore, W.L. Delano, P. Gros, R.W. Grosse-Kunstleve, J.S. Jiang, J. Kuszewski, M. Nilges, N.S. Pannu, R.J. Read, L.M. Rice, T. Simonson, and G.L. Warren. Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Cryst. D*, 54:905–921, 1998.

[41] T.A. Jones, J.Y. Zou, S.W. Cowan, and M. Kjeldgaard. Improved methods for building protein models in electron-density maps and the location of errors in these models. *Acta Cryst. A*, 47:110–119, 1991.

[42] G.M. Sheldrick and T.R. Schneider. Shelxl: High-resolution refinement. *Macromol. Cryst. B*, 277:319–343, 1997.

[43] D.M.F. van Aalten, W. Crielaard, K.J. Hellingwerf, and L. Joshua-Tor. Structure of the Photoactive Yellow Protein reconstituted with caffeic acid at 1.16 ångstrom resolution. *Acta Cryst. D*, 58:585–590, 2002.

[44] B.L. de Groot, D.M.F. van Aalten, R.M. Scheek, A. Amadei, G. Vriend, and H.J.C. Berendsen. Prediction of protein conformational freedom from distance constraints. *Proteins*, 29:240–251, 1997.

[45] A. Amadei, A.B.M. Linssen, and H.J.C. Berendsen. Essential dynamics of proteins. *Proteins*, 17:412–425, 1993.

[46] D.M.F. van Aalten, B.L. de Groot, J.B.C. Findlay, H.J.C. Berendsen, and A. Amadei. A comparison of techniques for calculating protein essential dynamics. *J. Comput. Chem.*, 18:169–181, 1997.

[47] L. Holm and C. Sander. Protein-structure comparison by alignment of distance matrices. *J. Mol. Biol.*, 233:123–138, 1993.

[48] T. Gensch, J. Hendriks, and K. J. Hellingwerf. Tryptophan fluorescence monitors structural changes accompanying signalling state formation in the photocycle of Photoactive Yellow Protein. *Photochem. Photobiol. Sci.*, 3:531–536, 2004.

[49] S. Crosson, S. Rajagopal, and K. Moffat. The LOV domain family: Photoresponsive signaling modules coupled to diverse output domains. *Biochemistry*, 42:2–10, 2003.

[50] T.E. Meyer, G. Tollin, J.H. Hazzard, and M.A. Cusanovich. Photoactive Yellow Protein from the purple phototrophic bacterium, *Ectothiorhodospira halophila* - Quantum yield of photobleaching and effects of temperature, alcohols, glycerol, and sucrose on kinetics of photobleaching and recovery. *Biophys. J.*, 56:559–564, 1989.

[51] J.J. van Beeumen, B.V. Vreese, S. M. van Bun, W.D. Hoff, K.J. Hellingwerf, T.E. Meyer, D.E. McCree, and M.A. Cusanovich. Primary structure of a Photoactive Yellow Protein from the phototrophic bacterium *Ectothiorhodospira halophila*, with evidence for the mass and the binding-site of the chromophore. *Protein Sci.*, 2:1114–1125, 1993.

[52] W.D. Hoff, I.H.M. van Stokkum, H.J. van Ramesdonk, M.E. van Brederode, A.M. Brouwer, J.C. Fitch, T.E. Meyer, R. van Grondelle, and K.J. Hellingwerf. Measurement and global analysis of the absorbency changes in the photocycle of the Photoactive Yellow Protein from *Ectothiorhodospira halophila*. *Biophys. J.*, 67:1691–1705, 1994.

[53] T.E. Meyer, M.A. Cusanovich, and G. Tollin. Transient proton uptake and release is associated with the photocycle of the Photoactive Yellow Protein from the purple phototrophic bacterium *Ectothiorhodospira halophila*. *Arch. Biochem. Biophys.*, 306:515–517, 1993.

[54] J. Hendriks, W.D. Hoff, W. Crielaard, and K.J. Hellingwerf. Protonation deprotonation reactions triggered by photoactivation of Photoactive Yellow Protein from *Ectothiorhodospira halophila*. *J. Biol. Chem.*, 274:17655–17660, 1999.

[55] J. Hendriks, I.H.M. van Stokkum, and K.J. Hellingwerf. Deuterium isotope effects in the photocycle transitions of the Photoactive Yellow Protein. *Biophys. J.*, 84:1180–1191, 2003.

[56] D.H. Pan, A. Philip, W.D. Hoff, and R.A. Mathies. Time-resolved resonance Raman structural studies of the pB′ intermediate in the photocycle of Photoactive Yellow Protein. *Biophys. J.*, 86:2374–2382, 2004.

[57] Z. Salamon, T.E. Meyer, and G. Tollin. Photobleaching of the Photoactive Yellow Protein from *Ectothiorhodospira halophila* promotes binding to lipid bilayers - Evidence from surface-plasmon resonance spectroscopy. *Biophys. J.*, 68:648–654, 1995.

[58] W.D. Hoff, A. Xie, I.H.M. van Stokkum, X.J. Tang, J. Gural, A.R. Kroon, and K.J. Hellingwerf. Global conformational changes upon receptor stimulation in Photoactive Yellow Protein. *Biochemistry*, 38:1009–1017, 1999.

[59] M.E. van Brederode, W.D. Hoff, I.H.M. van Stokkum, M.L. Groot, and K.J. Hellingwerf. Protein folding thermodynamics applied to the photocycle of the Photoactive Yellow Protein. *Biophys. J.*, 71:365–380, 1996.

[60] B.C. Lee, A. Pandit, P.A. Croonquist, and W.D. Hoff. Folding and signaling share the same pathway in a photoreceptor. *Proc. Nat. Ac. Sci. USA*, 98:9062–9067, 2001.

[61] E.F. Chen, T. Gensch, A.B. Gross, J. Hendriks, K.J. Hellingwerf, and D.S. Kliger. Dynamics of protein and chromophore structural changes in the photocycle of Photoactive Yellow Protein monitored by time-resolved optical rotatory dispersion. *Biochemistry*, 42:2062–2071, 2003.

[62] B.C. Lee, P.A. Croonquist, T.R. Sosnick, and W.D. Hoff. PAS domain receptor Photoactive Yellow Protein is converted to a molten globule state upon activation. *J. Biol. Chem.*, 276:20821–20823, 2001.

[63] T. Gensch, E.F. Chen, A.B. Gross, J.C. Hendriks, K.J. Hellingwerf, and D.S. Kliger. Dynamics of alteration of secondary structure in the photocycle of Photoactive Yellow Protein (PYP) as monitored by time-resolved optical rotary dispersion (TRORD). *Biophys. J.*, 82:314A–314A, 2002.

[64] K. Takeshita, Y. Imamoto, M. Kataoka, F. Tokunaga, and M. Terazima. Themodynamic and transport properties of intermediate states of the photocyclic reaction of Photoactive Yellow Protein. *Biochemistry*, 41:3037–3048, 2002.

[65] J. Hendriks, T. Gensch, L. Hviid, M.A. van der Horst, K.J. Hellingwerf, and J.J. van Thor. Transient exposure of hydrophobic surface in the Photoactive Yellow Protein monitored with Nile Red. *Biophys. J.*, 82:1632–1643, 2002.

[66] Y. Imamoto, H. Kamikubo, M. Harigai, N. Shimizu, and M. Kataoka. Light-induced global conformational change of Photoactive Yellow Protein in solution. *Biochemistry*, 41:13595–13601, 2002.

[67] M. Harigai, Y. Imamoto, H. Kamikubo, Y. Yamazaki, and M. Kataoka. Role of an N-terminal loop in the secondary structural change of Photoactive Yellow Protein. *Biochemistry*, 42:13893–13900, 2003.

[68] H. Kandori, T. Iwata, J. Hendriks, A. Maeda, and K. J. Hellingwerf. Water structural changes involved in the activation process of Photoactive Yellow Protein. *Biochemistry*, 39:7902–7909, 2000.

[69] V. Schmidt, R. Pahl, V. Srajer, S. Anderson, Z. Ren, H. Ihee, S. Rajagopal, and K. Moffat. Protein kinetics: Structures of intermediates and reaction mechanism from time-resolved X-ray data. *Proc. Nat. Ac. Sci. USA*, 101:4799–4804, 2004.

[70] Z. Ren, B. Perman, V. Srajer, T.Y. Teng, C. Pradervand, D. Bourgeois, F. Schotte, T. Ursby, R. Kort, M. Wulff, and K. Moffat. A molecular movie at 1.8 ångstrom resolution displays the photocycle of Photoactive Yellow Protein, a eubacterial blue-light receptor, from nanoseconds to seconds. *Biochemistry*, 40:13788–13801, 2001.

[71] C.J. Craven, N.M. Derix, J. Hendriks, R. Boelens, K.J. Hellingwerf, and R. Kaptein. Probing the nature of the blue-shifted intermediate of Photoactive Yellow Protein in solation by NMR: Hydrogen-deuterium exchange data and pH studies. *Biochemistry*, 39:14392–14399, 2000.

[72] N.M. Derix, R.W. Wechselberger, M.A. van der Horst, K.J. Hellingwerf, R. Boelens, R. Kaptein, and N.A.J. van Nuland. Lack of negative charge in the E46Q mutant of Photoactive Yellow Protein prevents partial unfolding of the blue-shifted intermediate. *Biochemistry*, 42:14501–14506, 2003.

[73] G. Groenhof, M. Bouxin-Cademartory, B. Hess, S.P. de Visser, H.J.C. Berendsen, M. Olivucci, A.E. Mark, and M.A. Robb. Photoactivation of the Photoactive Yellow Protein: Why photon absorption triggers a trans-to-cis isomerization of the chromophore in the protein. *J. Am. Chem. Soc.*, 126:4228–4233, 2004.

[74] G. Groenhof, M.F. Lensink, H.J.C. Berendsen, J.G. Snijders, and A.E. Mark. Signal transduction in the Photoactive Yellow Protein. I. Photon absorption and the isomerization of the chromophore. *Proteins*, 48:202–211, 2002.

[75] G. Groenhof, M.F. Lensink, H.J.C. Berendsen, and A.E. Mark. Signal transduction in the Photoactive Yellow Protein. II. Proton transfer initiates conformational changes. *Proteins*, 48:212–219, 2002.

[76] I. Antes, W. Thiel, and W.F. van Gunsteren. Molecular dynamics simulations of Photoactive Yellow Protein (PYP) in three states of its photocycle: A comparison with X-ray and NMR data and analysis of the effects of Glu46 deprotonation and mutation. *Eur. Biophys. J.*, 31:504–520, 2002.

[77] M. Shiozawa, M. Yoda, N. Kamiya, N. Asakawa, J. Higo, Y. Inoue, and M. Sakurai. Evidence for large structural fluctuations of the photobleached intermediate of Photoactive Yellow Protein in solution. *J. Am. Chem. Soc.*, 123:7445–7446, 2001.

[78] K. Itoh and M. Sasai. Dynamical transition and proteinquake in Photoative Yellow Protein. *Proc. Natl. Ac. Sci. USA*, 101:14736–14741, 2004.

[79] R. Swendsen and J. Wang. Replica Monte Carlo simulation of spin-glasses. *Phys. Rev. Lett.*, 57:2607–2609, 19860.

[80] E. Marinari and G. Parisi. Simulated tempering - A new Monte Carlo scheme. *Europhys. Lett.*, 19:451–458, 1992.

[81] Y. Sugita and Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, 314:141–151, 1999.

[82] D. Frenkel and B. Smit. *Understanding molecular simulation. From algorithms to applications.* Academic Press, San Diego, second edition, 2002.

[83] H. Nymeyer and A.E. García. Simulation of the folding equilibrium of α helical peptides: A comparison of the generalized born approximation with explicit solvent. *Proc. Natl. Ac. Sci. USA*, 100:13934–13939, 2003.

[84] R.H. Zhou. Exploring the protein folding free energy landscape: Coupling replica exchange method with P3ME/RESPA algorithm. *J. Mol. Graph. Mod.*, 22:451–463, 2004.

[85] A.E. García and K.Y. Sanbonmatsu. Exploring the energy landscape of a beta hairpin in explicit solvent. *Proteins*, 42:345–354, 2001.

[86] A.E. García and J.N. Onuchic. Folding a protein in a computer: An atomic description of the folding/unfolding of protein A. *Proc. Nat. Ac. Sci. USA*, 100:13898–13903, 2003.

[87] W.Y. Yang, J.W. Pitera, W.C. Swope, and M. Gruebele. Heterogeneous folding of the trpzip hairpin: Full atom simulation and experiment. *J. Mol. Biol.*, 336:241–251, 2004.

[88] P. Dux, G. Rubinstenn, G.W. Vuister, R. Boelens, F.A.A. Mulder, K. Hard, W.D. Hoff, A.R. Kroon, W. Crielaard, K.J. Hellingwerf, and R. Kaptein. Solution structure and backbone dynamics of the Photoactive Yellow Protein. *Biochemistry*, 37:12689–12699, 1998.

[89] R. Kort, K.J. Hellingwerf, and R.B.G. Ravelli. Initial events in the photocycle of Photoactive Yellow Protein. *J. Biol. Chem.*, 279:26417–26424, 2004.

[90] H.J.C. Berendsen, J.P.M. Postma, W.F. van Gunsteren, and J. Hermans. *Interaction models for water in relation to protein hydration*, pages 331–342. D. Reidel Publishing Company, Dordrecht, 1981.

[91] E. Lindahl, B. Hess, and D. van der Spoel. GROMACS 3.0: A package for molecular simulation and trajectory analysis. *J. Mol. Mod.*, 7:306–317, 2001.

[92] W.F. van Gunsteren and H.J.C. Berendsen. *Gromos-87 manual*. Biomos BV, Nijenborgh 4, 9747 AG Groningen, The Netherlands, 1987.

[93] A.R. van Buuren, S.J. Marrink, and H.J.C. Berendsen. A molecular dynamics study of the decane/water interface. *J. Phys. Chem.*, 97:9206–9212, 1993.

[94] A.E. Mark, S.P. van Helden, P.E. Smith, L.H.M. Janssen, and W.F. van Gunsteren. Convergence properties of free energy calculations: Cyclodextrin as a case study. *J. Am. Chem. Soc.*, 116:6293–6302, 1994.

[95] B. Hess, B. Bekker, H.J.C. Berendsen, and J.G.E.M. Fraaije. LINCS: A linear constraints solver for molecular simulations. *J. Comput. Chem*, 18:1463–1472, 1997.

[96] S. Miyamoto and P.A. Kollman. SETTLE: An analytical version of the SHAKE and the RATTLE algorithms for rigid water molecules. *J. Comput. Chem*, 13:952–962, 1997.

[97] H.J.C. Berendsen, J.P.M. Postma, W.F. van Gunsteren, A. DiNola, and J.R. Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81:3684–3690, 1984.

[98] W. Humphrey, A. Dalke, and K. Schulten. VMD – Visual Molecular Dynamics. *J. Mol. Graph.*, 14:33–38, 1996.

[99] H. Fan and A.E. Mark. Relative stability of protein structures determined by X-ray crystallography or NMR spectroscopy: A molecular dynamics simulation study. *proteins*, 53:111–120, 2004.

[100] S. Rajagopal, S. Anderson, V. Srajer, M. Schmidt, R. Pahl, and K. Moffat. A structural pathway for signaling in the E46Q mutant of Photoactive Yellow Protein. *Structure*, 1:1, 2004.

[101] S. Anderson, V. Srajer, R. Pahl, S. Rajagopal, F. Schotte, P. Anfinrud, M. Wulff, and K. Moffat. Chromophore conformation and the evolution of tertiary structural changes in Photoactive Yellow Protein. *Structure*, 12:1039–1045, 2004.

[102] C. Bernard, K. Houben, N. Derix, D. Marks, M. van der Horst, K. Hellingwerf, R. Boelens, R. Kaptein, and N. van Nuland. The solution structure of a transient photocycle intermediate: Δ25 Photoactive Yellow Protein. *Structure*, 13:953–962, 2005.

[103] P.G. Bolhuis. Transition-path sampling of beta-hairpin folding. *Proc. Natl. Acad. Sci.*, 100:12129–12134, 2003.

[104] J. Vreede, W. Crielaard, K.J. Hellingwerf, and P.G. Bolhuis. Predicting the signaling state of Photoactive Yellow Protein. *Biophys. J.*, 88:3525–3535, 2005.

[105] A.H. Xie, L. Kelemen, J. Hendriks, B.J. White, K.J. Hellingwerf, and W.D. Hoff. Formation of a new buried charge drives a large-amplitude protein quake in photoreceptor activation. *Biochemistry*, 40:1510–1517, 2001.

[106] W.D. Hoff, P. Dux, K. Hard, B. de Vreese, I.M. Nugteren-Roodzant, W. Crielaard, R. Boelens, R. Kaptein, J. van Beeumen, and K.J. Hellingwerf. Thiol ester-linked p-coumaric acid as a new photoactive prosthetic group in a protein with rhodopsin-like photochemistry. *Biochemistry*, 33:13959–13962, 1994.

[107] M.L. Groot, L. van Wilderen, D.S. Larsen, M.A. van der Horst, I.H. van Stokkum, K.J. Hellingwerf, and R. van Grondelle. Initial steps of signal generation in Photoactive Yellow Protein. *Biochemistry*, 42:10054–10059, 2003.

[108] S. Rajagopal, S. Anderson, H. Ihee, V. Srajer, M. Schmidt, R. Pahl, and K. Moffat. A structural pathway for signaling in Photoactive Yellow Protein. *Biophys. J.*, 86:83A–83A, 2004.

[109] J. Vreede, M.A. van der Horst, K.J. Hellingwerf, W. Crielaard, and D.M.F. van Aalten. PAS domains - Common structure and common flexibility. *J. Biol. Chem.*, 278:18434–18439, 2003.

[110] J. Hendriks, I.H. van Stokkum, W. Crielaard, and K.J. Hellingwerf. Kinetics of and intermediates in a photocycle branching reaction of the Photoactive Yellow Protein from Ectothiorhodospira halophila. *FEBS Lett.*, 458:252–256, 1999.

[111] P.G. Bolhuis. Kinetic pathways of beta-hairpin (un)folding in explicit solvent. *Biophys. J.*, 88:50–61, 2005.

[112] D. Frenkel. Speed-up of Monte Carlo simulations by sampling of rejected states. *Proc. Natl. Ac. Sci. USA*, 101:17571–17575, 2004.

[113] I. Coluzza and D. Frenkel. Virtual-move parallel tempering. *ChemPhysChem*, 6:1779–1783, 2005.

[114] T. Masciangioli, S. Devanathan, M.A. Cusanovich, G. Tollin, and M.A. El-Sayed. Probing the primary event in the photocycle of Photoactive Yellow Protein using photochemical hole-burning technique. *Photochem. Photobiol.*, 72:639–644, 2000.

[115] W.D. Hoff, S.L.S. Kwa, R. van Grondelle, and K.J. Hellingwerf. Low-temperature absorbency and fluorescence spectroscopy of the Photoactive Yellow Protein from *Ectothiorhodospira halophila*. *Photochem. Photobiol.*, 56:529–539, 1992.

[116] Y. Imamoto, M. Kataoka, and F. Tokunaga. Photoreaction cycle of photoactive yellow protein from *Ectothiorhodospira halophila* studied by low temperature spectroscopy. *Biochemistry*, 35:14047–14053, 1996.

[117] Y. Imamoto, H. Koshimizu, K. Mihara, O. Hisatomi, T. Mizukami, K. Tsujimoto, M. Kataoka, and F. Tokunaga. Roles of amino acid residues near the chromophore of Photoactive Yellow Protein. *Biochemistry*, 40:4679–4685, 2001.

[118] S. Anderson, S. Crosson, and K. Moffat. Short hydrogen bonds in Photoactive Yellow Protein. *Acta Cryst. D*, 60:1008–1016, 2004.

[119] H.M. Berman, T. Battistuz, T.N. Bhat, W.F. Bluhm, P.E. Bourne, K. Burkhardt, L. Iype, S. Jain, P. Fagan, J. Marvin, D. Padilla, V. Ravichandran, B. Schneider, N. Thanki, H. Weissig, J.D. Westbrook, and C. Zardecki. The protein data bank. *Acta Cryst. D*, 58:899–907, 2002.

[120] F.G. Parak. Proteins in action: The physics of structural fluctuations and conformational changes. *Curr. Op. Struct. Biol.*, 13:552–557, 2003.

[121] M.E. van Brederode, W.D. Hoff, I.H.M. van Stokkum, M.L. Groot, and K.J. Hellingwerf. Protein folding thermodynamics applied to the photocycle of the Photoactive Yellow Protein. *Biophys. J.*, 71:365–380, 1996.

[122] S. Reinelt, E. Hofmann, T. Gerharz, M. Bott, and D. R. Madden. The structure of the periplasmic ligand-binding domain of the sensor kinase CitA reveals the first extracellular PAS domain. *J. Biol. Chem.*, 278:39189–39196, 2003.

[123] A. Baltuska, I.H.M. van Stokkum, A. Kroon, R. Monshouwer, K.J. Hellingwerf, and R. van Grondelle. The primary events in the photoactivation of yellow protein. *Chem. Phys. Lett.*, 270:263–266, 1997.

[124] A. Pandini and L. Bonati. Conservation and specialization in PAS domain dynamics. *Protein Eng.*, 18:127–137, 2005.

[125] A.K. Smilde, R. Bro, and P. Geladi. *Multiway Analysis in Chemistry*. John Wiley, Chichester, 2004.

[126] L.R. Tucker. Some mathematical notes on three-factor analysis. *Psychometrika*, 31:110–182, 1966.

[127] C. de Ligny, M. Spanjer, J. van Houwelingen, and H. Weesie. Three-mode factor analysis of data on retention in normal-phase high performance liquid chromatography. *J. Chromatogr.*, 301:311–324, 1984.

[128] M. Parrinello and A. Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.*, 52:7182–7190, 1981.

[129] S. Nosé and M.L. Klein. Constant pressure molecular dynamics for molecular systems. *Mol. Phys.*, 50:1055–1076, 1983.

[130] S. Nosé. An extension of the canonical ensemble molecular dynamics method. *Mol. Phys.*, 52:187–191, 1986.

[131] W.G. Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A*, 31:1695–1697, 1985.

[132] H.J.C. Berendsen, D. van der Spoel, and R. van Drunen. GROMACS: A message-passing parallel molecular dynamics implementation. *Comp. Phys. Comm.*, 91:43–56, 1995.

[133] W.F. van Gunsteren, S.R. Billeter, A.A. Eising, P.H. Hünenberger, P. Krüger, A.E. Mark, W.R.P. Scott, and I.G. Tironi. *Biomolecular simulation: The GROMOS96 manual and user guide.* Vdf Hochschulverlag AG an der ETH Zürich, Zürich, Switzerland, 1996.

[134] X. Daura, A.E. Mark, and W.F. van Gunsteren. Parametrization of aliphatic CH_n united atoms of GROMOS96 force field. *J. Comput. Chem.*, 19:535–547, 1998.

[135] C.A. Andersson and R. Bro. The N-way toolbox for MATLAB. *Chemometr. Intell. Lab.*, 52:1–4, 2000. http://www.models.kvl.dk/source/nwaytoolbox/.

[136] P.J. Kraulis. MOLSCRIPT - A program to produce both detailed and schematic plots of protein structures. *J. Appl. Cryst.*, 24:946–950, 1991.

[137] E.A. Merritt and D.J. Bacon. Raster3D: Photorealistic molecular graphics. *Methods Enzymol.*, 277:505–524, 1997.

[138] W.C. Fuqua and S.C. Winans. A LuxR-LuxI type regulatory system activates agrobacterium Ti plasmid conjugal transfer in the presence of a plant tumor metabolite. *J. Bacteriol.*, 176:2796–2806, 1994.

[139] I.Y. Hwang, P.L. Li, L.H. Zhang, K.R. Piper, D.M. Cook, M.E. Tate, and S.K. Farrand. TraI, a LuxI homolog, is responsible for production of conjugation factor, the Ti plasmid n-acylhomoserine lactone autoinducer. *Proc. Natl. Ac. Sci. USA*, 91:4639–4643, 1994.

[140] Y.P. Qin, Z.Q. Luo, A.J. Smyth, P. Gao, S.B. von Bodman, and S.K. Farrand. Quorum-sensing signal binding results in dimerization of TraR and its release from membranes into the cytoplasm. *EMBO J.*, 19:5212–5221, 2000.

[141] M. Kataoka, S. Kosono, T. Seki, and T. Yoshida. Regulation of the transfer genes of streptomyces plasmid psn22 - In-vivo and in-vitro study of the interaction of TraR with promoter regions. *J. Bacteriol.*, 176:7291–7298, 1994.

[142] Y. Chai, J. Zhu, and S.C. Winans. TrlR, a defective TraR-like protein of *Agrobacterium tumefaciens*, blocks TraR function in vitro by forming inactive TrlR : TraR dimers. *Mol. Microbiol.*, 40:414–421, 2001.

[143] A. Vannini, C. Volpari, C. Gargioli, E. Muraglia, R. Cortese, R. de Francesco, P. Neddermann, and S. di Marco. The crystal structure of the quorum sensing protein trar bound to its autoinducer and target DNA. *EMBO Journal*, 21:4393–4401, 2002.

[144] R.G. Zhang, T. Pappas, J.L. Brace, P.C. Miller, T. Oulmassov, J.M. Molyneaux, J.C. Anderson, J.K. Bashkin, S.C. Winans, and A. Joachimiak. Structure of a bacterial quorum-sensing transcription factor complexed with pheromone and DNA. *Nature*, 417:971–974, 2002.

[145] Y. Yao, M.A. Martinez-Yamout, T.J. Dickerson, A.P. Brogan, P.E. Wright, and H.J. Dyson. Structure of the *Escherichia coli* quorum sensing protein SdiA: Activation of the folding switch by acyl homoserine lactones. *J. Mol. Biol.*, 262-273:355, 2006.

[146] J. Vreede, K.J. Hellingwerf, W. Crielaard, A. Smilde, and H.C.J. Hoefsloot. Protein triads, a new method to analyse protein dynamics. *Submitted to J. Comput. Biol.*, -:–, 2006.

[147] http://www.pymol.org.

[148] D.M.F. van Aalten, R. Bywater, J.B. Findlay, M. Hendlich, R.W. Hooft, and G. Vriend. PRODRG, a program for generating molecular topologies and unique molecular descriptors from coordinates of small molecules. *J. Comp. Aided Mol. Des.*, 10:255–262, 1996. http://davapc1.bioch.dundee.ac.uk/programs/prodrg/prodrg.html.

[149] C.E. White and S.C. Winans. Identification of amino acid residues of the *Agrobacterium tumefaciens* quorum-sensing regulator TraR that are critical for positive control of transcription. *Mol. Microbiol.*, 55:1473–1486, 2005.

[150] Z.Q. Luo, S.C. Su, and S.K. Farrand. In situ activation of the quorum-sensing transcription factor TraR by cognate and noncognate acyl-homoserine lactone ligands: Kinetics and consequences. *J. Bacteriol.*, 185:5665–5672, 2003.

[151] Y.R. Chai and S.C. Winans. Site-directed mutagenesis of a LuxR-type quorum-sensing transcription factor: Alteration of autoinducer specificity. *Mol. Microbiol.*, 51:765–776, 2004.

[152] Y.R. Chai and S.C. Winans. Amino-terminal protein fusions to the TraR quorum-sensing transcription factor enhance protein stability and autoinducer-independent activity. *J. Bacteriol.*, 187:1219–1226, 2005.

[153] M.H. Hefti, K.J. Francoijs, S.C. de Vries, R. Dixon, and J. Vervoort. The PAS fold - A redefinition of the PAS domain based upon structural prediction. *Eur. J. Biochem.*, 271:1198–1208, 2004.

[154] M.F. Perutz, M. Paoli, and A.M. Lesk. FixL, a haemoglobin that acts as an oxygen sensor: Signalling mechanism and structural basis of its homology with PAS domains. *Chem. Biol.*, 6:R291–R297, 1999.

[155] H. Kurokawa, D.S. Lee, M. Watanabe, I. Sagami, B. Mikami, C.S. Raman, and T. Shimizu. A redox-controlled molecular switch revealed by the crystal structure of a bacterial heme PAS sensor. *J. Biol. Chem.*, 279:20186–20193, 2004.

[156] H.J. Park, C. Suquet, J.D. Satterlee, and C.H. Kang. Insights into signal transduction involving PAS domain oxygen-sensing heme proteins from the X-ray crystal structure of *Escherichia coli* DOS heme domain (EcDosH). *Biochemistry*, 43:2738–2746, 2004.

[157] E.D. Getzoff, K.N. Gutwin, and U.K. Genick. Anticipatory active-site motions and chromophore distortion prime photoreceptor PYP for light activation. *Nature Struct. Biol.*, 10:663–668, 2003.

[158] S. Rajagopal and K. Moffat. Crystal structure of a Photoactive Yellow Protein from a sensor histidine kinase: Conformational variability and signal transduction. *Proc. Natl. Ac. Sci. USA*, 100:1649–1654, 2003.

[159] R. Fedorov, I. Slichting, E. Hartmann, T. Domratcheva, M. Fuhrmann, and P. Hegemann. Crystal structures and molecular mechanism of a light-induced signaling switch: The phot-LOV1 domain from *Chlamydomonas reinhardtii*. *Biophys. J.*, 84:2474–2482, 2003.

[160] C.A. Amezcua, S.M. Harper, J. Rutter, and K.H. Gardner. Structure and interactions of PAS kinase N-terminal PAS domain: Model for intramolecular kinase regulation. *Structure*, 10:1349–1361, 2002.

[161] P.B. Card, P.J.A. Erbel, and K.H. Gardner. Structural basis of arnt PAS-b dimerization: Use of a common beta-sheet interface for hetero- and homodimerization. *J. Mol. Biol.*, 353:664–677, 2005.

[162] P.J.A. Erbel, P.B. Card, O. Karakuzu, R.K. Bruick, and K.H. Gardner. Structural basis for PAS domain heterodimerization in the basic helix-loop-helix-PAS transcription factor hypoxia-inducible factor. *Proc. Natl. Ac. Sci. USA*, 100:15504–15509, 2003.

[163] O. Yildiz, M. Doi, I. Yujnovsky, L. Cardone, A. Berndt, S. Hennig, S. Schulze, C. Urbanke, P. Sassone-Corsi, and E. Wolf. Crystal structure and interactions of the PAS repeat region of the *Drosophila* clock protein PERIOD. *Mol. Cell*, 17:69–82, 2005.

[164] A. Razeto, V. Ramakrishnan, C.M. Litterst, K. Giller, C. Griesinger, T. Carlomagno, N. Lakomek, T. Heimburg, M. Lodrini, E. Pfitzner, and S. Becker. Structure of the NCOa-1/Src-1 PAS-b domain bound to the LXXLL motif of the STAT6 transactivation domain. *J. Mol. Biol.*, 336:319–329, 2004.

[165] J.R. Wagner, J.S. Brunzelle, K.T. Forest, and R.D. Vierstra. A light-sensing knot revealed by the structure of the chromophore-binding domain of phytochrome. *Nature*, 438:325–331, 2005.

[166] C.M. Dunham, E.M. Dioum, J.R. Tuckerman, G. Gonzalez, W.G. Scott, and M.A. Gilles-Gonzalez. A distal arginine in oxygen-sensing heme-PAS domains is essential to ligand binding, signal transduction, and structure. *Biochemistry*, 42:7701–7708, 2003.

[167] J. Key and K. Moffat. Crystal structures of deoxy and co-bound bjFixLh reveal details of ligand recognition and signaling. *Biochemistry*, 44:4627–4635, 2005.

[168] S.M. Harper, L.C. Neil, and K.H. Gardner. Structural basis of a phototropin light switch. *Science*, 301:1541–1544, 2003.

[169] S.M. Harper, J.M. Christie, and K.H. Gardner. Disruption of the LOV-J alpha helix interaction activates phototropin kinase activity. *Biochemistry*, 43:16184–16192, 2004.

[170] A. Chapman-Smith and M.L. Whitelaw. Novel DNA binding by a basic helix-loop-helix protein - The role of the dioxin receptor PAS domain. *J. Biol. Chem.*, 281:12535–12545, 2006.

[171] E.M. Dioum, J. Rutter, J.R. Tuckerman, G. Gonzalez, M.A. Gilles-Gonzales, and S.L. McKnight. NPAS2: A gas-responsive transcription factor. *Science*, 298:2385–2387, 2002.

[172] Y.S.J. Ho, L.M. Burden, and J.H. Hurley. Structure of the GAF domain, a ubiquitous signaling motif and a new class of cyclic GMP receptor. *EMBO J.*, 19:5288–5299, 2000.

[173] A. Fersht. *Protein Science*. W.H. Freeman and Company, New York, 1999.

[174] C. Levinthal. Are there pathways for protein folding? *J. Chim. Phys.*, 65:44–45, 1968.

[175] K.A. Dill and H.S. Chan. From Levinthal to pathways to funnels. *Nature Struct. Biol.*, 4:10–19, 1997.

[176] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, J. Anang, W. Miller, and D.J. Lipman. Gapped BLAST and PSI-BLAST: A new generation of protein based search programs. *Nucleic Acids Res.*, 25:3389–3402, 1997.

[177] R.D. Finn, J. Mistry, B. Schuster-Böckler, S. Griffiths-Jones, V. Hollich, T Lassman, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S.R. Eddy, E.L.L. Sonnhammer, and A. Bateman. Pfam: Clans, web tools and services. *Nucleic Acids Res.*, 34:D247–D251, 2006.

[178] D. van der Spoel, E. Lindahl, B. Hess, A.R. van Buuren, E. Apol, P.J. Meulenhoff, D.P. Tieleman, A.L.T.M. Sijbers, K.A. Feenstra, R. van Drunen, and H.J.C. Berendsen. *GROMACS User Manual, Version 3.3*. GROMACS, http://www.gromacs.org, 2005.

[179] Verlet L. Computer ”experiments” on classical fluids. *Phys. Rev.*, 195:98–103, 1967.

[180] R.W. Hockney, S.P. Goel, and J. Eastwood. Quiet high resolution computer models of a plasma. *J. Comput. Phys.*, 14:148–158, 1974.

[181] T. Darden, D. York, and L. Pedersen. Particle mesh Ewald: An N-log(N) method for Ewald sums in large systems. *J. Chem. Phys.*, 98:10089–10092, 1993.

[182] U. Essmann, L. Perera, M.L. Berkowitz, T. Darden, H. Lee, and L.G. Pedersen. A smooth particle mesh Ewald potential. *J. Chem. Phys.*, 103:8577–8592, 1995.

[183] H.C. Andersen. Molecular dynamics at constant pressure and/or temperature. *J. Chem. Phys.*, 72:2384–2393, 1980.

Summary

Cells interact with their environment and to survive, any cell has to continuously adapt to changes in external conditions. This requires the cell to perceive changes in its environment and subsequently generate an appropriate response to those changes, a process called signal transduction. As with many processes occurring in the living cell, the process of signal perception and signal transduction involve proteins, as described in *Chapter 1*. Proteins are chains of amino acids, folded in a specific three-dimensional shape. Via variation in the sequence of amino acids, each protein acquires a specific fold and specific properties uniquely suited to its biological function. Proteins with similar function and similar characteristics in amino acid sequence and structure are grouped in protein families. *Appendix A* gives a detailed overview of protein structure. Similarities in amino acid sequences have led to the discovery of the PAS family. PAS stands for PER, ARNT, SIM, the first three proteins exhibiting sequence similarities in combination with a biological function of signal transduction. When crystal structures of PAS domains became available, their structural resemblance was remarkable. The Photoactive Yellow Protein (PYP), a bacterial blue light sensor, was proposed as the structural prototype of the PAS family. This protein contains a covalently linked chromophore, enabling the protein to sense blue light.

The striking resemblance in the three-dimensional structures of PAS domains has led to the research presented in *Chapter 2*. The first part of the chapter describes the crystal structure of a minimal PAS fold. The design of the minimal PAS fold is based on PYP, via truncation of N-terminal residues that are not part of the PAS core. PYP maintains the PAS fold, even in absence of its N-terminal cap, and despite the exposure of several hydrophobic residues to solvent. The second part of Chapter 2 describes a comparison of structures in the PAS family, combining the structure of this Δ_{25}PYP and structures of other PAS domains. This has lead to the hypothesis that similar conformations might involve similar conformational changes, even though these domains bind a variety of ligands. Sampling the conformational space of PAS structures at atomic detail, followed by essential dynamics analysis of the C_α atoms of the common PAS core, shows that a segment, containing α-helices A and B, moves in a concerted fashion.

The main goal of the work presented in this thesis is to provide a better understanding of the mechanisms underlying signal perception and transduction in PAS domains. In PYP these processes have been the subject of many investigations, leading to a detailed description of the events taking place during its signal transduction cycle. Upon absorption of a blue light photon, *trans* to *cis* isomerisation of the chromophore occurs, followed by a proton transfer from a glutamate residue (Glu46) to the chromophore. Then, the protein partially unfolds, completing the formation of the signaling state within microseconds. The protein relaxes back to the resting state at a sub-second time scale. Due to the large range of time scales involved, some issues required advanced molecular dynamics techniques to understand the molecular mechanism underlying the light-triggered changes. The basics of molecular dynamics simulations are described in *Appendix B*. In *Chapter 3* parallel tempering simulations were used to predict the mechanism of formation and the structure of the signaling state of PYP. The proton transfer from Glu46 to the chromophore results in a transfer of negative charge from the chromophore to Glu46. At Glu46 this charge is localized over a smaller set of atoms in comparison to the chromophore, leading to destabilization of the hydrogen bond network inside the protein, allowing water molecules to access the protein interior and loosening the interaction between the chromophore and the protein. The next step is the solvent exposure of Glu46, followed by the solvent exposure of the chromophore. The solvent exposure of Glu46 has also an effect on the stability of the hydrophobic core in the N-terminal domain, resulting in the partial unfolding seen in several experiments.

The prediction of Chapter 3 was confirmed by the NMR structure of Δ_{25}-PYP in its signaling state, that also showed a solvent exposed chromophore and Glu46. Further comparison revealed however, that the NMR structure has a more extended conformation for the central β-sheet and a larger degree of unfolding in the α-helical region containing Glu46. *Chapter 4* presents a parallel tempering simulation of the truncated mutant protein showing that solvent exposure of the chromophore and Glu46 occurs more easily with respect to the wild type. Moreover, the partially unfolded conformations compare well with the NMR structure. These observations further substantiate the prediction of Chapter 3 and provide insight into the role of the N-terminal domain. By restricting the conformational freedom of the central β-sheet, the N-terminal cap limits the pathways for and the extent of light-triggered unfolding.

Chapters 3 and 4 also discuss the conformational characteristics of the resting state, using parallel tempering simulations. These show that the chromophore is in contact with solvent molecules, even in the resting state.

Chapter 4 ends with an investigation of the recovery of the resting state, a huge challenge for the current molecular simulation techniques, due to the time scales involved. Starting from a partially unfolded representative conformation of the signaling state, Chapter 4 presents parallel tempering simulations initiated from different chromophore configurations. Complete refolding did not occur in any of the recovery simulations. Nevertheless, formation of a crucial hydrogen bond in the protein interior was observed, as well as incorporation of the chromophore in the protein core. The simulations show that the *cis* or *trans* configuration of the chromophore has a significant effect on the ability of the protein to refold into its receptor state. A simulation with a lowered rotational barrier of the chromophore double bond further substantiated this conclusion.

The Essential Dynamics analyses performed in Chapter 2 construct a model for protein motions by summarizing the data, *i.e.* an ensemble of protein structures, in the temporal direction and in atomic fluctuations. When arranging the data from a molecular dynamics simulation as a three-way array, the application of three-way analysis techniques, such as Tucker-3 is possible. This method uses three sets of dynamically relevant components that describe (i) the relative fluctuation of the atoms, (ii) the time profile of a motion and (iii) the direction of a motion respectively. *Chapter 5* presents the application of Tucker-3 analysis to protein dynamics: Protein Triads. Both Essential Dynamics and Protein Triads are applied to a simple system of eleven atoms moving in a hinge-bending motion, and the parallel tempering simulation of the formation of the signaling state of PYP, as discussed in Chapter 3. Both methods perform equally well in capturing the variance in the data, but Protein Triads is able to separate the relevant motion into different model components, resulting in a better interpretable model. For the simulation of PYP, this means that the N-terminal cap is the most flexible part in the protein, exhibiting motions that are linked to the functional unfolding of the chromophore binding pocket, and flexibilities inherent to the N-terminal region. Chapter 5 ends with a comparison of simulations of three PAS domains. Using Essential Dynamics for such a comparison requires the different protein structural ensembles to have identical size. To achieve this, protein parts not belonging to the PAS definition were omitted in the analysis. Using Protein Triads, such a comparison is facilitated by the scalar product of the directional modes, and does not require the definition of a common core.

Chapters 2-5 all investigate PAS domains in isolation. *Chapter 6* describes molecular dynamics simulations of TraR, a quorum sensor from *Agrobacterium tumefaciens*. The Tra quorum sensing system represents a one-component signaling pathway with a PAS domain as input domain. Bacteria use chemical signals to regulate gene expression at high population densities. These cell-cell communication systems require the release and detection of molecules that can diffuse through the plasma membrane. Receptor proteins sense the concentration of these molecules: When this concentration is above a certain threshold level activation of a specific set of genes occurs. After sensing, the receptors activate or repress transcription of genes. These genes may code for proteins involved in various functions, including pathogenesis, sporulation and gene transfer. This process is referred to as quorum sensing. TraR is functional as a dimer, with the two monomers in different conformations, one compact and one elongated. The simulations presented in Chapter 6 reveal that the auto-inducer and its close environment show conformational heterogeneity, involving several water molecules: It is an interplay between the auto-inducer and water molecules in its proximity that induces the conformational variety. Moreover, the auto-inducer also affects residues not in its direct environment. The DNA binding domain shows significantly enhanced fluctuations in presence of the auto-inducer in the elongated conformation. In other words, the presence of the auto-inducer in the PAS domain induces fluctuations in the HTH motif that are instrumental for DNA binding. Simulations at high temperature of the compact monomer show that the PAS domain and the DNA binding domain move away from each other, possibly facilitating conversion of the compact monomer to the elongated conformation. Summarizing these observations results in a mechanism for the DNA binding of TraR and possibly its sensitivity to proteolysis. The auto-inducers in the TraR dimer induce increased fluctuations in their close surroundings. As a consequence, the HTH motifs dangle from the PAS scaffold, searching for DNA. Binding of DNA induces the formation of an interface between the two HTH motifs, and is accomodated by the formation of the more compact m1 conformer.

Since the discovery of the PAS family, knowledge on the PAS fold has expanded widely. *Chapter 7* gives an overview of the current literature on PAS domains. Many routes for signal transduction have been identified for PAS domains, depending on the location of the co-factor and the location of the contact site with the signal transduction interaction partner(s). Despite these differences, the central β-sheet in the PAS fold is a stable scaffold.

Samenvatting

Elke cel staat in contact met zijn omgeving. Om te overleven moet een cel zich steeds aanpassen aan de veranderingen in zijn externe omstandigheden. Dit is alleen maar mogelijk als de cel veranderingen in zijn omgeving kan waarnemen, en vervolgens een geschikte reactie op deze veranderingen kan genereren. Dit proces staat bekend als signaaloverdracht. Bijna alle processen in de cel worden uitgevoerd door eiwitten, en zo ook het signaaloverdrachtsproces, zoals beschreven in *Hoofdstuk 1*. Eiwitten bestaan uit een keten van aminozuren, die gevouwen is in een specifieke driedimensionale vorm. Door variatie in de volgorde van de aminozuren heeft elk eiwit een specifieke vorm en eigenschappen, waardoor het in staat is zijn biologische functie uit te voeren. *Appendix A* geeft een gedetailleerd overzicht van de structuur en opbouw van eiwitten. Op basis van functie en overeenkomsten in aminozuurvolgorde kunnen eiwitten ingedeeld worden in eiwitfamilies. Op deze manier is de PAS familie geïdentificeerd. PAS staat voor PER, ARNT, SIM, de eerste drie eiwitten met gelijkende aminozuursequentie en signaaloverdracht als biologische functie. Toen er kristalstructuren van leden van deze familie beschikbaar kwamen, bleek dat er opmerkelijke overeenkomsten waren in driedimensionale structuur. Het Photoactive Yellow Protein (PYP, foto-actief geel eiwit), een bacteriële sensor voor blauw licht, is voorgesteld als prototype voor de PAS familie. Dit eiwit bevat een covalent gebonden chromofoor, waardoor het eiwit blauw licht kan waarnemen.

De opvallende gelijkenis in de driedimensionale structuren van PAS domeinen heeft geleid tot het onderzoek beschreven in *Hoofdstuk 2*. Het eerste deel van het hoofdstuk beschrijft de kristalstructuur van een minimale PAS vouwing. Het ontwerp van de minimale PAS vouwing is gebaseerd op een afgekapte mutant van PYP waarin de N-terminale aminozuren die geen deel uitmaken van de PAS kern zijn verwijderd. Zelfs in afwezigheid van deze N-terminale kap behoudt PYP de PAS vouwing, ondanks de blootstelling van verscheidene hydrofobe groepen aan water. Het tweede gedeelte van Hoofdstuk 2 beschrijft een vergelijking van structuren in de PAS familie, inclusief de structuur van de Δ_{25}-PYP. Dit heeft geleid tot the hypothese dat vergelijkbare structuren tot vergelijkbare conformationele veranderingen kunnen leiden, ook al binden de domeinen verschillende soorten liganden. Het in kaart brengen van de conformationele ruimte van PAS structuren in atomair detail, gevolgd door een essentïele dynamica analyse van de C_α atomen van de gemeenschappelijke PAS kern, laat zien dat een segment, met α-helices A en B, een collectieve beweging uitvoert.

Het hoofddoel van het werk beschreven in dit proefschrift is het verkrijgen van beter begrip van de mechanismen betrokken bij het waarnemen en vervolgens doorgeven van een signaal. In PYP zijn deze processen het onderwerp geweest van vele onderzoeken, die hebben geleid tot een gedetailleerde beschrijving van de gebeurtenissen tijdens de signaaloverdrachtscyclus.

Direct na absorptie van een blauw licht foton isomeriseert de chromofoor, gevolgd door een protonoverdracht van glutamaat (Glu46) naar de chromofoor. Dan ontvouwt het eiwit gedeeltelijk om zo de vorming van de signaaltoestand compleet te maken op een tijdsschaal van microseconden. Binnen een seconde relaxeert het eiwit relaxeert terug naar de rustende toestand. Vanwege de grote spreiding in tijdsschalen hebben sommige onderdelen geavanceerde moleculaire simulatiemethoden nodig om de moleculaire mechanismen van de lichtgestuurde processen te begrijpen. De basis van moleculaire dynamica simulaties is beschreven in *Appendix B*. In *Hoofdstuk 3* zijn parallel tempering simulaties gebruikt om het mechanisme van de vorming van de signaaltoestand van PYP te voorspellen. Dit proces begint met de protonoverdracht van Glu46 naar de chromofoor, waardoor de negatieve lading van de chromofoor naar Glu46 verplaatst. Op het glutamaat is de negatieve lading minder goed gestabiliseerd, resulterend in destabilisatie van het waterstofbrugnetwerk binnen het eiwit. Dit heeft tot gevolg dat water moleculen de chromofoorbindingsplaats binnen kunnen komen en de interactie tussen het eiwit en de chromofoor losser maken. Glu46 en de chromofoor gaan vervolgens naar de buitenkant van het eiwit, waar beide groepen blootgesteld zijn aan bulk water. De nieuwe locatie van Glu46 beïnvloedt de stabiliteit van de hydrofobe kern van de N-terminale domein, resulterend in de gedeeltelijke ontvouwing gezien in verschillende experimenten.

De voorspelling van Hoofdstuk 3 is later bevestigd door de NMR structuur van de mutant Δ_{25}-PYP in de signaaltoestand, met eveneens de chromofoor en Glu46 buiten het eiwit. Verdere vergelijking laat zien dat β-sheet in de NMR structuur meer uitgestrekt is en dat de regio rijk aan α-helices, en met Glu46, meer ontvouwen is. *Hoofdstuk 4* presenteert een parallel tempering simulatie van de afgekapte mutant die laat zien dat de blootstelling van de chromofoor en Glu46 gemakkelijker verloopt in vergelijking tot het wild type. De gedeeltelijk ontvouwen structuren lijken veel op de NMR structuur. Deze waarnemingen bevestigen de voorspelling van hoofdstuk 3 en geven inzicht in de rol van het N-terminale domein. Door het beperken van de bewegingsvrijheid van de centrale β-sheet beperkt de N-terminale kap de paden en de uitgestrektheid van de licht-gestuurde ontvouwing. Hoofdstukken 3 en 4 bespreken ook de conformationele eigenschappen van de rustende toestand, resulterend uit parallel tempering simulaties. Deze laten zien dat de chromofoor in contact is met oplosmiddel moleculen.

Hoofdstuk 4 eindigt met een onderzoek naar de terugreactie van de rustende toestand van PYP. Dit is een enorme uitdaging voor de huidige simulatiemethoden, vanwege de lange duur: de terugreactie is een proces op subseconde tijdsschaal. Uitgaand van een gedeeltelijk ontvouwen conformatie van de signaaltoestand, met verschillende chromofoorconformaties, bespreekt Hoofdstuk 4 parallel tempering simulaties van de terugreactie. In geen enkele terugreactiesimulatie vond volledige terugvouwing van het eiwit plaats. Desalniettemin is de vorming van een cruciale waterstofbrug waargenomen, evenals het binnengaan van het eiwit door de chromofoor. De simulaties laten zien dat de *cis* of de *trans* configuratie een significant effect heeft op hoe goed het eiwit in staat is terug te vouwen naar de receptor toestand. Een simulatie van een verlaagde rotatiebarriere geeft een verdere onderbouwing van deze conclusie.

De essentiële dynamics analyses in hoofdstuk 2 bestaan uit een model voor eiwitbewegingen verkregen door de data, bestaand uit een ensemble van eiwitstructuren, samen te vatten als temporele en atomaire fluctuaties. Wanneer de data van een moleculaire dynamica simulatie als een drieweg matrix gerangschikt wordt, is de toepassing van drieweg analyse methoden mogelijk, zoals Tucker-3. Deze methode maakt gebruik van drie sets van dynamisch relevante componenten die (i) de relatieve fluctuatie van de atomen (ii) het tijdsprofiel van een beweging en (iii) de richting van een beweging beschrijven. *Chapter 5* presenteert de toepassing van Tucker-3 analyse op eiwitdynamica: Protein Triads. Zowel essentiële dynamica als Protein Triads worden toegepast op een eenvoudig systeem van elf atomen die een scharnierbeweging uitvoeren en op een parallel tempering simulatie van de vorming van de signaaltoestand van PYP, beschreven in hoofdstuk 3. Beide methoden doen het even goed wat betreft het beschrijven van de variantie van de data. Protein Triads kan bovendien de relevante bewegingen van elkaar scheiden in de verschillende componenten van het model, resulterend in een beter interpreteerbaar model. Voor de simulatie van PYP betekent dit dat de N-terminale kap het meest flexibele gedeelte van het eiwit is, en dat de bewegingen van dit domein gelinkt zijn aan de functionele ontvouwing van de chromofoorbindingsplaats, maar ook typisch zijn voor de N-terminale regio. Hoofdstuk 5 eindigt met een vergelijking van simulaties van drie PAS domeinen. Een dergelijke vergelijking met essentiële dynamica vereist de selectie van een gelijk aantal atomen in de verschillende eiwitten. Voor een aantal structuren van de PAS familie werden die gedeeltes die geen onderdeel uitmaakten van de PAS kern niet meegenomen in de analyse. Met Protein Triads is een dergelijke vergelijking mogelijk door het inporduct van de richtingscomponenten te nemen, en is het niet nodig om een gemeenschappelijke kern te definiëren.

Hoofdstukken 2-5 beschrijven onderzoek van PAS domeinen in isolement. *Chapter 6* beschrijft moleculaire dynamica simulaties van TraR, een quorum sensor van *Agrobacterium tumefaciens*. Het Tra quorum waarnemingssysteem is een voorbeeld van een een-component signaleringspad met een PAS domein als input domein. Bacteriën gebruiken chemische signalen om genexpressie bij een hoge populatiedichtheid te reguleren. Deze communicatie van cel naar cel vereist de verspreiding en waarneming van moleculen, auto-inducers, die door het membraan heen kunnen diffunderen. Receptoreiwitten nemen de concentratie waar van deze moleculen en bij het bereiken van een bepaalde concentratie worden specifieke genen geactiveerd. Na het waarnemen activeren of verhinderen deze receptoren de transcriptie van genen. Deze genen kunnen coderen voor verschillende soorten eiwitten, met als functie pathogenesis, sporulatie en genoverdracht. Dit proces staat bekend als quorum waarneming. TraR functioneert als een dimeer, met de twee monomeren in verschillende conformaties, eén compact en de ander uitgestrekt. De simulaties beschreven in hoofdstuk 6 laten zien dat de auto-inducer en nabije omgevingverschillende conformaties kan aanemen. Het is een samenspel van de auto-inducer met watermoleculen in zijn nabijheid die de conformationele verscheidenheid veroorzaakt. De auto-inducer beïnvloedt ook verder verwijderde groepen. Het DNA bindend gedeelte beweegt significant meer als de auto-inducer aanwezig is, in de uitgestrekte conformatie. Anders gezegd, de auto-inducer in het PAS domein induceert fluctuaties in het HTH motief die instrumenteel zijn voor DNA binding. Hoge temperatuursimulaties laten zien dat het PAS domein en de DNA bindende motief van elkaar af bewegen in de compacte monomeer. Dit is mogelijk een omzetting van de compacte naar de uitgestrekte vorm. samenvattend, deze waarnemingen resulteren in het formuleren van een mechanisme voor het binden van DNA door TraR en verklaren mogelijk de gevoeligheid voor proteolyse. De auto-inducers veroorzaken meer fluctuaties in hun directe omgeving, en als gevolg daarvan laten ze de DNA bindende motieven bengelen, zoekend naar DNA. Het binden van DNA vormt een tussenvlak tussen de twee HTH motieven, en wordt geaccomodeerd door het vormen van de compacte monomeer.

Sinds de eerste definitie van de PAS familie is de kennis over de PAS vouwing flink toegenomen.

Chapter 7 geeft een overzicht van de huidige literatuur over PAS domeinen. Er zijn inmiddels verschillende routes voor signaaloverdracht door PAS domeinen geïdentificeerd, afhankelijk van de locatie van de cofqactor en de contactplaats met de signaaloverdracht interactie partners. Ondanks deze verschillen is de centrale β-sheet een stabiel structuurelement in de PAS structuur.

Curriculum vitae

Jocelyne Vreede was born on February 17th in the year 1978 in Geleen, The Netherlands. At the age of 18 she graduated at the gymnasium of Scholengemeenschap Sint Michiel in Geleen and went to Amsterdam to study Chemistry at the University of Amsterdam. In her third year she followed a course in molecular simulation and since then, the field of molecular simulation has held her interest. She performed her Master's project in the molecular simulation group of Berend Smit, investigating the efficiency of surfactants in separating oil-water mixtures, using a coarse-grained soft-spheres model. In March 2001 she obtained her Master of Science degree and in May, the same year, she started on her PhD in the microbiology group of Klaas Hellingwerf.

Jocelyne was involved in teaching activities during her PhD project, including Bioinformatics courses for Biology and Chemistry students. Also she supervised the Master's theses of Michiel van Lun, Gerda Conijn and Yan Jiang. In December 2004 and April 2005 she was a guest teacher at the Hogeschool Leiden, giving courses on protein bioinformatics.

She is married to Steven Dinkelaar and mother of Lianne, who was born on October 1st in 2005.

Publications

- J. Vreede, W. Crielaard, K.J. Hellingwerf and P.G. Bolhuis *Predicting the signaling state of Photoactive Yellow Protein* Biophys. J. 2005, 88:3525-3535
- J. Vreede, M.A. van der Horst, R. Kort, W. Crielaard and K.J. Hellingwerf *Measuring and modelling dynamical changes in the structure of Photoactive Yellow Protein* Phase Transitions 2004, 77:3-20
- J. Vreede, M.A. van der Horst, K.J. Hellingwerf, W. Crielaard and D.M.F. van Aalten *PAS domains - Common structure and common flexibility?* J. Biol. Chem. 2003, 278:18434-18439
- L. Rekvig, M. Kranenburg, J. Vreede, B. Hafskjold and B. Smit*Investigation of surfactant efficiency using Dissipative Particle Dynamics* Langmuir 2003, 19:8195-8205

Manuscripts in submission and preparation

- J. Vreede, K.J. Hellingwerf, A.K. Smilde, W. Crielaard and H.C.J. Hoefsloot *Protein Triads: A new method to analyse collective motions in proteins* Submitted to J. Comput. Biol.
- E.J.M. Leenders, L. Guidoni, U. Röthlisberger, J. Vreede, P.G. Bolhuis and E.J. Meijer *Protonation of the chromophore in Photoactive Yellow Protein* Submitted to J. Phys. Chem. B.
- J. Vreede, K.J. Hellingwerf and W. Crielaard *Auto-inducer mediates TraR DNA binding through backbone fluctuations* Submitted to FEBS Letters
- J. Vreede, K.J. Hellingwerf and P.G. Bolhuis *Conformational requirements for efficient recovery of the pG state of Photoactive Yellow Protein* In preparation

Nawoord

Het is klaar!!

Het begon met mijn huisgenoot Gabrielle. Haar vader, Klaas Hellingwerf, zocht iemand die simulaties van eiwitten kon doen. Ik vond mezelf wel geschikt en ging naar het kantoor van Klaas. Nadat ik me had voorgesteld als huisgenoot van zijn dochter, en vertelde geïnteresseerd te zijn in een promotieplaats, werd ik uitgenodigd om plaats te nemen en me te onderwerpen aan een kruisverhoor. Ik was hier totaal niet op voorbereid, maar desalniettemin beantwoordde ik alle vragen van Klaas zo goed als ik kon. Zijn laatste vraag was "Wanneer wil je beginnen?".

Dat werd dus 1 mei 2001. Mijn eerste ontmoeting met Klaas was typerend voor onze verdere omgang. Ik kon altijd terecht, maar dan kreeg ik ook lastige vragen. Eén keer, tijdens mijn zwangerschap, kreeg ik echter geen lastige vragen, maar een verlenging.

Mijn directe begeleider Wim Crielaard was heel anders. Tijdens boswandelingen in Lunteren en lange autoritten naar de synchrotron in Hamburg bespraken we uitgebreid de projecten waaraan ik werkte. We spraken ook veel over carrière maken met een gezin. Wim is voor mij een voorbeeld hoe je als werkende ouders een gelukkig gezin kunt zijn.

Wim en Klaas zijn microbiologen die mij, met mijn moleculaire simulaties, soms moeilijk konden helpen. In het begin heb ik veel geleerd van Daan van Aalten in Dundee, die me eiwitkristallografie en Essential Dynamics heeft bijgebracht. Ondanks pogingen van Daan me over te halen kristallografie te gaan doen, heeft mijn werk tijdens de promotie vooral bestaan uit simulaties. Dit is voor een deel te danken aan Peter Bolhuis. Uren hebben we samen naar filmpjes van het Photoactive Yellow Protein zitten kijken. De congressen biofysica in de USA die we samen bezochten waren erg interessant en gezellig. De scripts van Jarek, werkzaam in Peters groep, hebben enorm geholpen bij het doen van de parallel tempering simulaties.

Samen met Wim ging ik vaak langs bij Huub Hoefsloot en Age Smilde, om te praten over het analyseren van bewegingen in eiwitten. Deze bijeenkomsten bestonden voornamelijk uit elkaar proberen te begrijpen.

Bij microbiologie was ik een buitenbeentje. Terwijl mijn collega's in het lab labdingen deden, was ik aan het rekenen en programmeren. Toch was ik niet geïsoleerd: ik heb veel gehad aan discussies met Johnny, Sergey, Remco, Michael en Martijn. De spelletjes Catan met Ania, Alex en Femke waren gezellig, en goed voor het kwijtraken van frustraties ("It's all pointless!").

Mijn simulaties heb ik voornamelijk uitgevoerd met GROMACS. Via de GROMACS mailing list leerde ik Gerrit Groenhof kennen. Een lange tijd hebben we vrijwel dagelijks emailcontact gehad, met onderwerpen als PYP, promotiefrustraties, werk zoeken, kinderen krijgen en kinderen hebben.

Ik deed de simulaties op de supercomputers van SARA en op de beowulfclusters van de Molsim groep, luisterend naar goede muziek op ITS-radio (/radio1).

Veel nieuwe en/of goede ideeën voor onderzoek kreeg ik tijdens het sporten. Gelukkig waren er vrienden om mee te hardlopen, roeien, zwemmen, klimmen, squashen, en snowboarden en kon ik vaak collega's ook nog zo gek krijgen om mee te doen.

De afgelopen paar jaar was ik afwisselend vrolijk, treurig en af en toe heel boos, afhankelijk van hoe het onderzoek liep. Deze stemmingswisselingen werden erger tijdens de laatste maanden van schrijven. Lieve vrienden en (stief/schoon)familie, bedankt voor jullie steun en geduld! Pappa, bedankt voor de juiste woorden op het juiste moment. Mamma en Cecile, heel fijn dat jullie altijd wilden luisteren naar mijn gezeur.

Steven, zonder jouw kookkunsten na een lange werkdag, je mooie motiverende preken tijdens mijn promotiedipjes, je gezichtsuitdrukkingen vol onbegrip als ik probeerde mijn onderzoek uit te leggen, je sterke uithuilschouders en, ja, zonder jou, had ik het nooit voor elkaar gekregen.

En nu?

Nou, dat lijkt me duidelijk: Lianne.