Thermodynamic driving forces in protein regulation studied by molecular dynamics simulations

Dissertation
Zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultäten
der Georg-August-Universität zu Göttingen

Vorgelegt von
Ulf Hensen
aus Aachen

Göttingen, 2008

D7
Referent: Prof. Dr. M. Suhm
Korreferent: Prof. Dr. H. Grubmüller
Tag der mündlichen Prüfung: 22.1.2009

Contents

Chapter 1

Introduction

Proteins are biological linear heteropolymeric macromolecules composed from unbranched chains of amino acids. From a chemical point of view, proteins are a rather homogeneous class of molecules. However, despite the relatively low number of 20 different monomeric units, they are characterized by a seemingly endless variability of structure and function. In fact, their versatility stands strikingly out in comparison with other types of biologically important molecules such as carbohydrates, lipids, or nucleic acids making them nature's choice for performing most difficult duties in every single living cell on Earth. For example, they fulfill tasks as different as cell statics (building blocks in coats of viruses, in epidermal keratin, and collagen in bones and cartilage), transport and storage of other (macro-)molecules and electrons, signalling between cells and organs (hormones), support of the immune system (antibodies), conversion of chemical into mechanical energy (in muscles), control of cell membrane thoroughfares for other molecules and ions, regulation of gene expression, regulation of gene transcription, and chaperoning other proteins whilst they obtain their folded structure. To describe this versatility, Lesk [1] noted that *"In the drama of life on a molecular scale, proteins are where the action is."*

Accordingly, for a whole lot of scientific disciplines, proteins are fascinating objects to study, both from academic and practical considerations. For example, material science is interested in the extraordinary combination of stability and flexibility of spider silk or collagen proteins, engineering strives to imitate properties of proteins as molecular nanomotors, and medicine is interested in proteins due to the fact that protein malfunction causes a considerable number of diseases, e.g. mutations are the reason for some types of cancer, protein aggregation is responsible for neurodegenerative disorders including Huntington, Alzheimer, Creutzfeld-Jacob, or motor neuron disease. Understanding

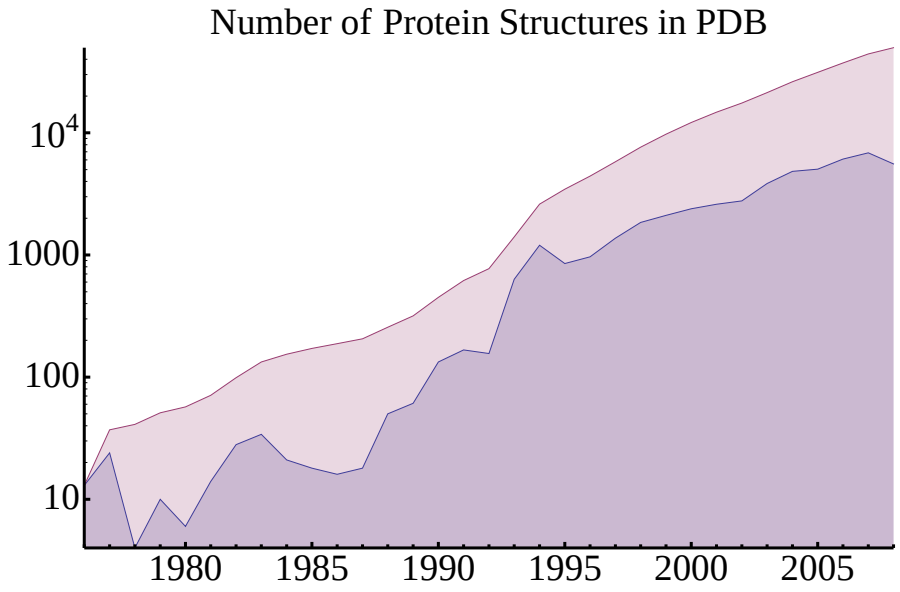

Figure 1.1: Exponential growth of the three-dimensional structures available in the Protein Data Base (PDB). Purple: Total number of entries; blue: number of yearly added entries (source: pdb.org)

protein function is also essential for rational drug design aiming at finding means of interrupting essential protein function of pathogens such as HIV, SARS, Hepatitis, etc., or at the bacterial ribosome.

Apart from these practical considerations, it is interesting to learn about how such versatile molecular machines as proteins work, and why. This interest is reflected by the exponential growth of solved spatial structures of proteins deposited in the Protein Data Base (PDB). Starting with 13 structures in 1976, the number of entries will reach 50000 this year (see Fig. 1.1) doubling the number of records approximately every three years.

The structural information deposited there is a prerequisite for the understanding of protein function, albeit not always a sufficient one. In many cases protein function is inheritly linked to protein dynamics, apart from obvious cases where proteins are mere scaffolds for cells. For many other cases, protein function is impeded when protein motion is restricted. For example, observations of the temperature dependence of crystallographic B-factors indicate predominantly restricted motion below temperatures where enzyme activity is abolished, the so-called glass transition temperature [2].

Compared to structural information, our knowledge of protein dynamics is still rather

fragmentary. To a large extent this is due to the fact that the most common method for structure determination – X-ray diffraction [3, 4] – provides no direct information on dynamics, as B-factors and missing electron density do not allow to distinguish between static and dynamic disorder. The other common method for protein structure determination, nuclear magnetic resonance (NMR) [5, 6], likewise mostly assumes the molecules to be rigid bodies. Time-resolved X-ray diffraction [7, 8] allows picosecond dynamic resolution, but is, due to the tremendous experimental complexity, as yet rarely used. Time-resolved NMR techniques have also made remarkable progress and are based on measurement of relaxation parameters [9–11], chemical shift averaging and backbone proton exchange rates, which can be made independently of structure determination and give direct information on protein motion. A variety of spectroscopic methods that can also monitor protein dynamics, albeit not in atomic detail, include neutron scattering [12, 13], fluorescence transfer and depolarization, depolarized light scattering, electron paramagnetic resonance (EPR), resonance Raman and Mössbauer spectroscopy and time-resolved Fourier transform infrared spectroscopy (trFTIR) [14–16]. At the single-molecule level, the atomic force microscopy-based "mechanical triangulation" has been recently shown to measure sub-nanometer interatomic distances [17].

With the exception of time-resolved X-ray diffraction, those experiments are often difficult to interpret at the atomic level. Here, atomistic computer simulations contribute to the detailed understanding of molecular mechanisms. In this thesis, classical molecular dynamics (MD, for details see Chapter 2) has been used, an atomistic simulation method, which treats each atom as a point mass and describes the interaction between atoms by simple force terms. Trajectories are obtained by integrating Newton's equations of motion. The large number of pair-wise interactions to be evaluated and the short time steps required by the fastest motions (such as bond and angle vibrations) lead to very high computational effort. This currently limits MD to system sizes of about a million atoms and to timescales of several hundred nanoseconds.

Most relevant biological processes, such as gating of ion channels, allosteric signalling, ligand binding, molecular recognition, etc., in contrast, occur at longer timescales and are thus, in principle, out of reach of current MD simulations. This is true despite numerous efforts to speed up calculations, i.e. multistep methods [18–23], fast multipole methods [24–27], and Ewald summation techniques [28]. Also, the use of constraints is popular to allow for larger step sizes [29–31]. Processes beyond the microsecond limit are still only accessible by special enhanced sampling techniques, which, unfortunately, result in loss of

dynamical information and often of thermodynamic accuracy as well [32]. This results in, among others, difficulties concerning compare simulation results with experimental data. One simulation technique, force-probe molecular dynamics (FPMD) [33] stands out in this respect, since it allows to enforce configurational rearrangement of proteins and at the same time closely mimic the corresponding experimental setups, namely atomic force microscopy (AFM) [34–36]. This technique was used in Chapter 3 and will be discussed in some more detail in the methods section Chapter 2. If the reaction coordinate is not a priori clear (like it is in, e.g., muscle proteins), however, FPMD is difficult to apply.

This is the case in allosterically regulated proteins. Allosteric regulation is widely used in biological systems as an effective mechanism to control protein activity [37–40] by transmitting information across long distances within the protein. Allostery is so called because it involves the binding of a ligand at one site of the protein causing a change of affinity or catalytic efficiency at a far distant site. While allostery has been subject to extensive research efforts for more than 40 years, the mechanisms underlying the communication between distant sites and energetically couple them remain largely elusive [39, 41]. The classical view tends to explain allostery in terms of discrete configurational rearrangements that eventually result in a different mean structure [42, 43]. In a thermodynamic framework, this view emphasizes the enthalpic contributions (i.e. change of mean structure), whereas it neglects changes in dynamic fluctuations around this average structure (i.e. entropic contributions). In contrast, allosteric processes could, in principle, proceed through changes in protein dynamics without any change of mean structure whatsoever [44]. Recent experiments for the catabolite activator protein (CAP^N) have suggested such an exclusively entropic effect [41] but could not be verified by subsequent studies [45]. Most allosteric proteins probably follow a mixture of these two effects, but a clear physical picture is lacking so far.

To a large extent this is due to the fact that most studies have focused on the comparison between liganded and unliganded end states. However, no single structural view will reveal an allosteric mechanism, since without information on intermediate states no examination of the allosteric pathways can be made. Accordingly, the intermediate states are what must be in the main focus of examination. These, however, cannot be isolated in positively cooperative proteins, since their singly liganded states are too poorly populated. In contrast, negatively cooperative systems allow, in principle, for the characterization of intermediate states [46, 47], but data on those is as yet even rarer than for the positively cooperative systems. Computer simulations are in principle apt to significantly contribute

to the field since, here, intermediate states can be selectively focused on.

Two main points, however, have to be considered major obstacles for successful application of computer simulations in this field. First, allosteric proteins are typically oligomers with high molecular weights, meaning that the number of particles to be simulated is large and the computational expense is similarly large. Second, the timescale at which allosteric regulation works is typically on the order of milliseconds to seconds and, thus, far out of reach of standard MD.

Aims of this thesis

The principal interest of this work is to explore the driving forces behind protein regulation by means of molecular dynamics simulations. In Chapter 3, the force-induced activation of titin kinase (TK) is examined by means of force-probe MD. The simulation results are linked to experimental atomic force microscopy results and provide a distinct picture of how force exerted on TK by contraction of the muscle it is embedded in results in enzymatic activation. As a prerequisite for the accurate description of the entropic contributions of allosteric regulation processes, in Chapters 4–6, a new method for the estimation of configurational entropy is developed. Eventually, in Chapter 7, the allosteric regulation of *Leishmania mexicana* pyruvate kinase is examined.

Force-induced activation of titin kinase

The giant muscle protein titin is one of the three main parts of the basic contractile machinery of muscles, the sarcomere. It is so called because of its size of approximately 3 MDa, the largest protein known to date. It spans half of the sarcomere from the Z-disk to the M-band connecting the other two parts of the sarcomere, actin and myosin, and guaranteeing the resting elasticity of the muscle [48, 49] by realigning the muscle filaments and restoring the sarcomere length after muscle contraction. On the M-band side of titin, a serine/threonine protein kinase domain (TK) is located, the only catalytic domain of titin. This catalytic domain is unusual in the sense that, in contrast to most other kinases, titin kinase is autoinhibited by its C-terminal regulatory tail which blocks the active site. TK activation, thus, must be preceded by release of autoinhibition. From previous molecular dynamics simulations, it is known that force exerted on TK might induce release of autoinhibition [50]. Despite further evidence by subsequent experiments [51, 52], definite proof of mechanical activation rather than mere partial unfolding upon force

exertion was hitherto lacking. Chapter 3 presents results of a combined experimental and theoretical effort to elucidate the mechanisms underlying TK regulation. The focus here is on the molecular dynamics simulations, which helped to interprete the accompanying single-molecule force spectroscopy and enzymatic experiments.

The independent subspace approach for the estimation of the absolute configurational entropy of macromolecules

Both entropies S and Gibbs free energies G (or A for Helmholtz free energies, depending on the ensemble) are fundamental quantities in statistical mechanics with eminent importance in biological processes. While G is a criterion of directionality in all chemical processes, in the field of structural biology it is essential for studying of, e.g., protein-protein and protein-ligand interactions, enzymatic reactions, or transport through membranes. S is a measure of order and the main driving force behind many of the biologically relevant phenomena including protein folding and other processes driven by hydrophobic forces. Both quantities are difficult to obtain from computer simulations and, therefore, considerable attention has been devoted in the past 50 years to developing new methods addressing this issue [53**?** –55]. In many cases, however, either efficiency or accuracy or both are unsatisfactory. In particular, the computation of configurational entropies which is needed for, e.g., the description of protein regulation, mostly relies on the quasi-harmonic approximation [56, 57], which does not include proper description of non-linear and higher-order correlations. In Chapters 4–6, a new method is developed which accurately includes those crucial correlations into the description of the configurational entropies of macromolecules.

Allosteric regulation of L. mexicana pyruvate kinase

The allosterically regulated pyruvate kinase (PYK) it a large 2000-residue homotetrameric enzyme which catalyzes the final step of the glycolysis transforming phosphoenol pyruvate (PEP) to pyruvate. The activity of this enzyme shows positive cooperativity both homotropically (i.e., upon binding of the substrate PEP itself) and heterotropically upon binding the allosteric activator fructose-bisphosphate. Despite extensive enzymatic, crystallographic and mutagenesis studies, the mechanism of allosteric PYK regulation is largely elusive. It is not known, how binding of an allosteric activator triggers changes in neighbouring subunits' catalytic sites more than 10 nm away. *l.mexicana* PYK is the only

isoform which has been crystallised in both inactive T and active R-state. As elaborated in Chapter 7, we examined the dynamics underlying the allosteric PYK regulation at the atomic level using extensive MD simulations in explicit solvent. The large size of this enzyme (the simulation system with explicit solvent comprised about $500,000$ atoms) and the fact that allosteric signals are typically very subtle rendered this project computationally quite challenging.

Chapter 2
2 Theory, Methods, and Concepts

This chapter sketches the methodological framework of this thesis. Subdivided into seven parts it comprises summaries of classical molecular dynamics simulations, enforced molecular dynamics, free energy calculations, and dimension reduction techniques such as principal component analyis (PCA). It is not intended to exhaustively account for all details, but rather to provide sufficient background to understand the results. Fore more detailed accounts, see recent reviews [58] and relevant text books [59, 60].

2.1 Molecular dynamics

Molecular dynamics (MD) simulations rely on numerical integrations of Newton's equations of motions in order to describe the time evolution of a molecular system. It is the only established method which is hitherto able to accurately account for both the high complexity of proteins consisting of up to a million atoms and the relatively long time scales underlying their functional dynamics. This is achieved by incorporating three drastic approximations to the exact time-dependent Schrödinger equation which are described below.

Born-Oppenheimer Approximation

An exact description of the dynamics of any atomic system is provided by the time-dependent Schrödinger equation

$$i\hbar\frac{\partial\psi}{\partial t} = \mathcal{H}\psi\,,$$

where $\mathcal{H}$ is the Hamilton operator, i.e. the sum of kinetic and potential energy operators, ψ is the wavefunction, and $\hbar = h/2\pi$ with Planck's constant h. The wave function ψ is a function of the positions of both nuclei and electrons. The Born-Oppenheimer approximation [61] aims at separating the fast electronic degrees of freedom from the slow nuclear ones, i.e. it assumes that the electrons instantaneously follow the slow nuclear motion. The electronic wave function ψ_e, consequently, depends only parametrically on the nuclear coordinates, and the total wave function can be separated into an electronic and a nuclear part,

$$\psi_{tot}(\mathbf{R}, \mathbf{r}) = \psi_n(\mathbf{R})\psi_e(\mathbf{R}; \mathbf{r}),$$

where $\mathbf{R} = (\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_N)$ and $\mathbf{r} = (\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_K)$ denote the coordinates of the N nuclei and K electrons, respectively. The Schrödinger equation then splits into two parts, a time-dependent Schrödinger equation of the motion of the nuclei, and a time-independent Schrödinger equation for the electronic problem,

$$\mathcal{H}_e\psi_e = E_e\psi_e,$$

where $\mathcal{H}_e$ is the electronic Hamiltonian and E_e is the ground state energy, which parametrically depends on the nuclear positions $\mathbf{R}$. Solving this time-independent part, one obtains the potential energy hyper surface $V(\mathbf{R})$, the potential of the fast electrons in which the slow nuclei move.

Classical Dynamics

The motion of the nuclei in this potential energy surface $V(\mathbf{R})$ is described by the time-dependent Schrödinger equation for the nuclear dynamics, which, however, cannot be computed for a macromolecular system with thousands of atoms. Classical MD assumes that Newton's equations of motion provide a sufficiently accurate description

$$\begin{aligned} m_i \frac{d^2\mathbf{R}_i(t)}{dt^2} &= -\nabla_i V(\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_N), \text{ or} \\ m_i\mathbf{a}_i &= \mathbf{F}_i, \end{aligned}$$

where $\mathbf{R}_i$ and m_i are position and mass of nucleus i, respectively. The force $\mathbf{F}_i$ acting on nucleus i determines its acceleration $\mathbf{a}_i$, which in turn results in a change of position and velocity within a given discrete time step Δt. This time step has to be chosen small

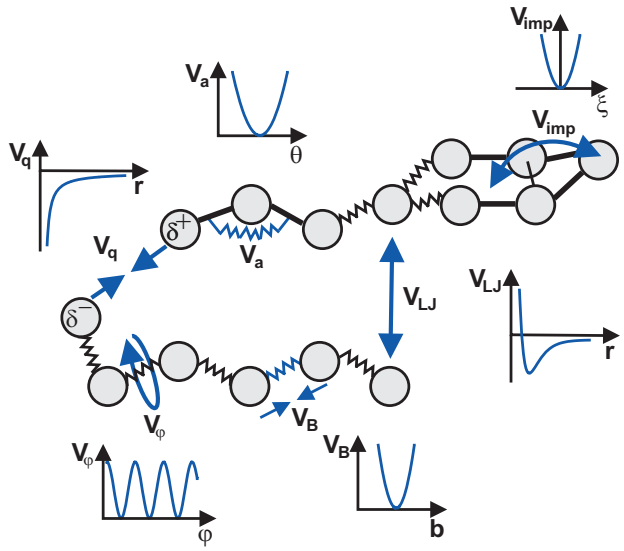

Figure 2.1: Illustration of the components of a typical molecular mechanics force field as well as a sketch of their functional form as explained in Eq. 2.1.

compared to the fastest motions in the system. Bond and angle vibrations involving hydrogen atoms occur at a femtosecond timescale restricting the time step to about 1 fs. A number of algorithms allowing larger time steps by constraining bond length has been devised. In this work, the LINCS algorithm [31] and a time step of 2 fs has been chosen. Newtons's equations of motion were computed via the leap-frog modification of the Verlet algorithm [62],

$$\mathbf{R}(t+\Delta t) = \mathbf{R}(t) + \mathbf{v}\left(t+\frac{1}{2}\Delta t\right)\Delta t$$

$$\mathbf{v}\left(t+\frac{1}{2}\Delta t\right) = \mathbf{v}\left(t-\frac{1}{2}\Delta t\right) + \mathbf{a}(t)\Delta t,$$

where the current position $\mathbf{R}(t)$ and the accelerations $\mathbf{a}(t)$ are stored together with the mid-step velocities $\mathbf{v}(t-\Delta t/2)$.

Force Fields

The two above approximations are the basis of the Quantum mechanics molecular dynamics methods [59], which obtain the potential $V(\mathbf{R})$ and its gradient by solving the time-independent Schrödinger equation for the electronic degrees of freedom. For the

large number of electrons present in proteins this is impossible. A third approximation is required to address this problem. The potential $V(\mathbf{R})$ is approximated as a sum of simple energy terms

$$\begin{aligned} V &= \sum_{\text{bonds i}} V_b^i + \sum_{\text{bond angles j}} V_a^i + \sum_{\text{impropers k}} V_{\text{imp}}^k \\ &\quad + \sum_{\text{dihedrals l}} V_D^l + \sum_{\text{pairs } \alpha,\beta} \left(V_q^{\alpha,\beta} + V_{\text{LJ}}^{\alpha,\beta} \right), \end{aligned} \tag{2.1}$$

where the bonded energy terms are defined to be either harmonic (V_b, V_a, V_{imp}) or cosine functions (V_D), and the non-bonded terms are based on physical laws (V_q, V_{LJ}),respectively. Accordingly,

$$\begin{aligned} V &= \sum_{\text{bonds i}} \frac{k_i}{2} \left(l_i - l_{i,0}\right)^2 + \sum_{\text{bond angles j}} \frac{k_j}{2} \left(\theta_j - \theta_{j,0}\right)^2 \\ &\quad + \sum_{\text{impropers k}} \frac{k_k}{2} \left(\zeta_k - \zeta_{k,0}\right)^2 \\ &\quad + \sum_{\text{dihedrals l}} \frac{V_l}{2} \left(1 + \cos\left(n\omega - \gamma\right)\right) \\ &\quad + \sum_{\text{pairs}\alpha,\beta} \frac{q_\alpha q_\beta}{4\pi\epsilon_0\epsilon_1 \mathbf{R}_{\alpha,\beta}} + 4\epsilon_{\alpha,\beta} \left[\frac{\sigma_{\alpha,\beta}^{12}}{R_{\alpha,\beta}^{12}} - \frac{\sigma_{\alpha,\beta}^{6}}{R_{\alpha,\beta}^{6}} \right]. \end{aligned}$$

Force-field parameters for the bonded interactions include the equilibrium bond length $l_{i,0}$, bond angles $\theta_{j,0}$, force constants k, as well as multiplicity n, torsional barrier height V_l and phase γ of the dihedral angles l. Non-bonded interactions are parameterized in terms of partial charges q_α for Coulomb interactions as well as the repulsion and attraction coefficients, σ^{12} and σ^6, of the Lennard-Jones potential which represents van-der-Waals attraction and Pauli repulsion, respectively. For computational efficiency, many-body effects are usually not explicitly considered, which entails quadratic scaling of the computational cost with the number of particles.

Different force field have been developed, which differ in the number of terms in Eq. 2.1, the exact functional form as well as the parameters. In this thesis, the OPLS force field [63, 64] was used for biomolecules. Different versions of GROMOS [65] united atom force

fields were employed for some test cases described in Chapters 4–6. While the former treats all atoms explicitly, the latter merge aliphatic, non-polar hydrogen atoms with the carbons they are bound to. The resulting CH, CH_2 and CH_3 groups are treated as a single particle, with properly adjusted mass, charge, and Lennard-Jones parameters. Other popular force fields include AMBER [66], CHARMM [67] and MM3 [68]. All aforementioned force fields have in common that they have been developed for proteins on the basis of their smallest subunits, amino acids.

Parameters incorporated into a particular force field are usually obtained set-wise by iteratively fitting to experimental data and quantum-chemical calculations until the whole set of parameters fulfills the optimization requirements. This means that any single parameter differs slightly from its original value. Consequently, including new, hitherto unrepresented molecules into the force-field can be a cumbersome task, because in principle the force field as a whole would have to be reconsidered. Yet this is usually not done since the disadvantages in form of time-consumption by far outweight the benefits. Accordingly, the force field parameters needed for ligands in Chapters 3 and 7 were obtained by adopting similar molecules' parameters complemented by quantum chemical calculations without reparametrization of the whole force field.

Simulation details

Periodic boundaries All protein simulations in this thesis were simulated in explicit water. To avoid artifacts stemming from boundaries such as evaporation, surface tension, or preferred orientations of the molecules close to the boundaries, periodic boundaries were used in all cases. Periodic boundaries are achieved by infinitely replicating the simulation box and, for the force calculations, including interactions of each particle into account only with its peers within the 'mother-box' or the nearest-image. This way the simulation box does not have any surface. However, new artifacts can emerge because the macromolecules can also interact with their mirror image due to long-range electrostatic interactions. These periodicity artifacts can be minimized by choosing the box size large enough. Different box shapes, e.g. cubic, triclinic, truncated octahedron or dodecahedron, allow to optimally fit proteins inside thereby keeping protein-protein distances sufficiently large whilst also minimizing the number of solvent molecules. In this thesis, box sizes were chosen such that the minimal protein-border distance was 1 nm.

Thermostats In principle, solving the Newtonian equations conserves the total energy of the system and one, consequently, obtains a NVE ensemble. However, a molecular system of the size typically studied with MD simulations inevitably exchanges energy with its surroundings. This energy exchange is introduced into the simulation by connecting to a thermostat, i.e. coupling the system to a heat bath of reference temperature T_0. Here, the very robust Berendsen thermostat [69] has been employed which introduces temperature conservation by rescaling the velocities in every step with

$$v' = v\sqrt{1 + \frac{\Delta t}{\tau_T}\left(\frac{T_0}{T} - 1\right)},$$

where Δt is the integration time step and τ_T is the coupling constant defining the coupling strength. If not noted otherwise, this was set to 0.1 ps.

Barostats Biological systems are mostly subjected to constant pressure rather than constant volume, i.e. NPT ensembles are required. In addition to temperature conservation, therefore, pressure coupling was also introduced in form of the Berendsen barostat [69] which rescales coordinates in every step. Modifying the box size slightly in every step it thereby guarantees constant pressure thoughout the simulation.

Computational efficiency Further means purely for the sake of increased computational efficiency have been applied; most important are constraint algorithms (e.g., LINCS [31] for hydrogen bonds and SETTLE [30] for water models) and special handling of non-bonded interactions. These are the most expensive part of the force calculations. Even though multi-body interactions were not explicitly considered during the computations, the sum of the last term in Eq. 2.1 requires N^2 single computations for the N atoms of the system. A very straightforward way of reducing this number is ignoring all neighbors beyond a certain cut-off, typically in the range of 1.0 to 1.4 nm [67]. While this is justified for the fast decaying Lennard-Jones potential (which decreases with r^{-6}) severe artefacts have been reported for the slowly (r^{-1}) decaying Coulomb interactions [70].

The grid-based Ewald summation [71] overcomes this problem by splitting the electrostatics into two parts. The part within the cut-off is computed directly, and the long-range interactions outside the cut-off are computed in reciprocal space. Further improvement is achieved by the Particle Mesh Ewald (PME) summation, which employs fast Fourier transformations for the computation of the reciprocal sum [72] and scales

with $N \log N$. PME has been used wherever possible thoughout the works in this thesis.

2.2 Stochastic dynamics

Stochastic dynamics adds friction and a noise term to the Newtonian equations

$$m_i \frac{d^2\mathbf{R}_i(t)}{dt^2} = -m_i \xi_i \frac{d\mathbf{R}_i(t)}{dt} + \mathbf{F}_i\left(\mathbf{R}\right) + \mathring{\mathbf{R}}_i(t),$$

where ξ_i is a friction constant and the stochastic random force $\mathring{\mathbf{R}}_i(t)$ is assumed to be a stationary Gaussian random variable with zero mean and to have no correlation with prior velocities or with the systematic force, i.e. $\langle \mathring{\mathbf{R}}_i(t) \mathring{\mathbf{R}}_j(t + \Delta t) \rangle = 2 m_i \xi_i k_\mathrm{B} T \delta_{ij} \delta(t)$. Here, $\langle \dots \rangle$ denotes the ensemble average, k_B is the Boltzmann constant and δ is the Dirac delta function. Stochastic dynamics is particularily useful for simulations of systems in vacuum to mimic the forces exerted by the solvent on the molecule. In this thesis, stochastic dynamics has been used for the generation of thermodynamic integration (TI) reference ensembles of small test systems whose configurational entropies were subsequently computed with the method developed in Chapter 5.

2.3 Force-probe MD

Atomic Force Microscopy (AFM)

Atomic force microscopy (AFM) is widely used as a single-molecule experiment. A typical experimental setup is illustrated in Fig. 2.2. Here, the molecule under investigation is fixed between the surface on one side and an AFM-tip attached to a flexible cantilever on the other side. Moving the surface downwards exerts a pulling force on the examined molecule which is measured by the cantilever as a function of the surface displacement. The resulting force profile is read as a fingerprint of the mechanical stability of the molecule from which unfolding pathways to be inferred. As a single-molecule experiment, it provides information on single events rather than merely ensemble averaged values obtained from bulk experiments. In this respect it is easier to relate MD simulation results to AFM experiments than it is to bulk experiments, since in these, single molecule information is not available. The downside of AFM, however, is the fact that atomic interpretation of the observed force profiles is not readily at hand. Here, computer simulations can constitute useful complements.

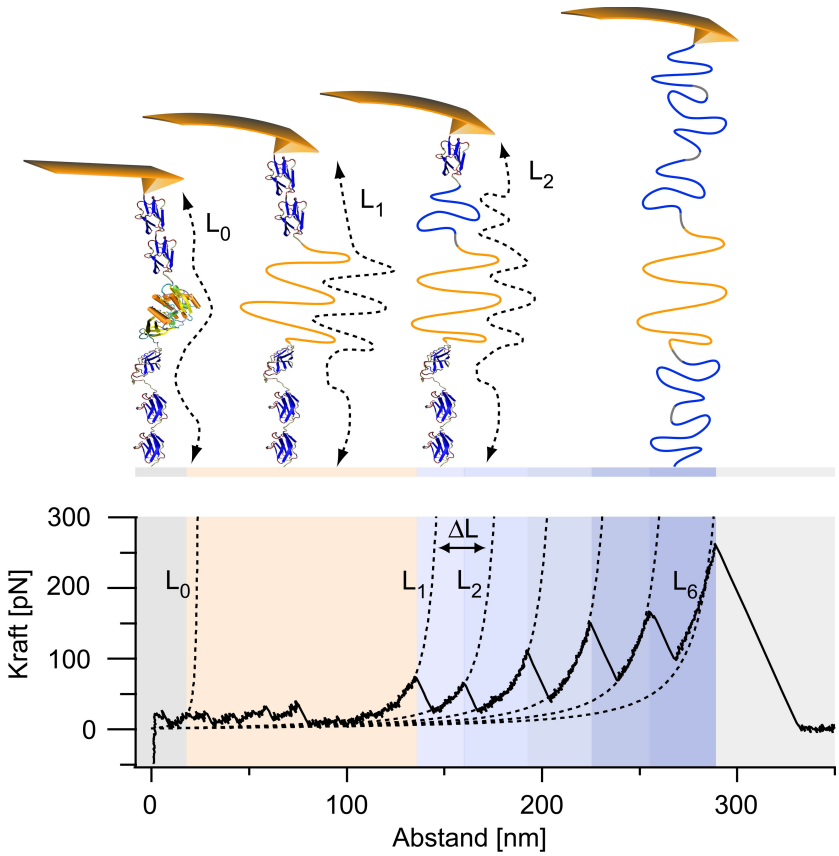

Figure 2.2: Illustration of a typical atomic force microscopy (AFM) setup [73]: Top: A molecule (here titin kinase, which was examined in detail in Chapter 3) is attached via a linker to the surface on one side and to a cantilever with AFM tip on the other side. The surface is then slowly moved downwards exerting a force on the examined molecule. The cantilever measures this force as a function of the surface displacement. Bottom: A saw-tooth pattern is typically recorded for the unfolding of proteins. The abrupt drops following the maxima correspond to rupture/unfolding forces. The contour L_i length of each unfolding event i is obtained by fitting each peak separately with a worm-like chain model [74].

Simulation setup

With classical MD, protein dynamics can now be routinely examined provided the process of interest occurs on the nano- to microsecond timescale or faster. Many biological phenomena studied in this thesis, such as allosteric protein regulation or force-induced unfolding, are considerably slower and thus out of reach for standard MD. Force-probe MD (FPMD) [75] circumvents this problem by applying external forces to parts of the protein. This way, the system is driven along a pre-defined pulling coordinate and energy barriers are more easily overcome (cf. Fig. 2.3). One or more atoms i are being subjected to a harmonic spring potential,

$$V_{\mathrm{spring},i}(t) = k_0 \left(z_i(t) - z_{\mathrm{spring},i}(t)\right)^2 ,$$

where k_0 is the force constant of the spring, $z_i(t)$ the position of the centre-of-mass (COM) of the pulled atoms, and $z_{\mathrm{spring},i}$ the position of the spring. The spring is then moved with constant velocity v in pulling direction $z_{\mathrm{spring},i}(t) = z_i(0) + vt$. The pulled atoms are then subject to an additional force,

$$F_i = k_0 \left(z_i(t) - z_{\mathrm{spring},i}(t)\right) .$$

FPMD is particularily apt to complement experiments which exert forces on single molecules such as atomic force microscopy (AFM) experiments discussed above or magnetic/optical tweezers. It has been shown to be a valuable tool for the interpretation of experimental data [76]. Among other cases, it has helped to explain questions of carbohydrate stiffness [77], ligand unbinding from proteins [75], or partial force-induced protein unfolding [50]. In Chapter 3, the force-induced unfolding of titin kinase is examined with this technique and compared to AFM experiments.

2.4 Essential dynamics

In cases where the reaction coordinate is a priori unclear, force probe MD is not an appropriate option to enforce slow processes. In those cases, such as allosteric regulation (7), we resorted to essential dynamics sampling (ED sampling). In ED simulations, the system is being driven along a reaction coordinate defined by one or more collective modes observed during a free dynamics simulation. These collective modes, or essential modes,

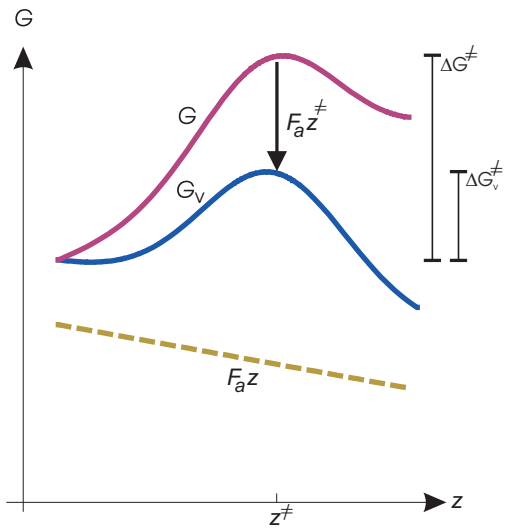

Figure 2.3: Decrease of transition barriers by AFM or FPMD. The additional force F_{ext} accelerates a transition along the reaction coordinate z (which is at the same time the pulling direction) by lowering the free energy barrier from $\Delta G^{\neq}$ to $\Delta G_v^{\neq}$.

are usually being obtained by Principal Component Analysis (PCA). PCA diagonalizes the covariance and obtains linearily uncorrelated eigenvectors plus corresponding eigenvalue, which indicates the amplitude of the collective motion along that particular eigenvector. PCA is discussed in more detail in Section 2.6.

Fig. 2.4 illustrates the basic idea of ED sampling. A set of eigenvectors is defined as essential subspace where the position of the system shall be manipulated, i.e. enhanced sampling is desired. Each of these eigenvectors can be subject to one of the available algorithms along the selected eigenvectors (7.10). The system is then required to fulfill distance constraints along the essential modes assigned to a particular algorithm. Modes orthogonal to the essential subspace are left completely unperturbed and can equilibrate during the simulation. In Chapter 7, essential dynamics was used as a relatively soft way of driving the allosteric enzyme pyruvate kinase from inactive to active configurations.

2.5 Free energy calculations

Differences of the free energies ΔG and entropies ΔS of two states A and B are commonly calculated by thermodynamic integration (TI) over physical quantities such as energy,

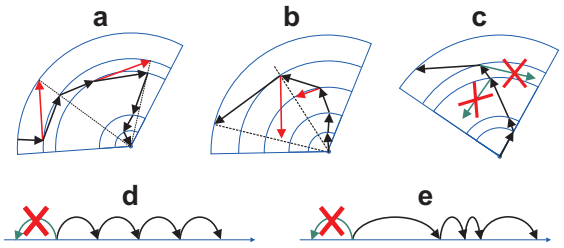

Figure 2.4: Principles of Essential Dynamics sampling. Black arrows: accepted steps, red: corrected steps, green: rejected steps. a, radius contraction. The distance from the target (centre circle) is required not to increase. Steps away from the target are rejected (red) and set to a constant radius position as illustrated with the dashed lines; b: fixed linear expansion, *radfix*. Starting from the centre, the radius has to increase by a predefined distance in each step. Steps not fulfilling the distance constraints (red arrows) are set according to the dashed lines; c: Acceptance radius expansion, *radacc*. Like *radfix*, except that the increment size is not fixed and steps into the wrong direction are merely rejected and the step is started again; d: linear expansion along a predefined direction (to the right) with predefined step size, *linfix*. Steps in the wrong direction (green) are being rejected (crossed); e: same as d, except that the step size is variable.

temperature or specific heat, or alchemical parameters.

In principle, both free energy differences and absolute free energies can be obtained; the latter requires a reference state B whose absolute free energy is known, such as an ideal gas. For this thesis, this was done using a parameter λ to simulate the transition between small molecules (state A with $\lambda = 0$) and non-interacting particles in harmonic wells (state B with $\lambda = 1$), whose absolute free energy was known. The free energy difference is then obtained according to

$$\Delta G = \int_0^1 \left\langle \frac{\partial H\left(\mathbf{R}, \lambda\right)}{\partial \lambda} \right\rangle_\lambda d\lambda,$$

where the total Hamiltonian of the system at any particular λ-value is $H = (1 - \lambda) H_B + \lambda H_A$. Here, one assumes the simulations at any intermediate λ to be in equilibrium. This can either be achieved by using the slow growth method [54] with sufficiently slow transition time or, as it was done in this thesis, discrete TI with system at N different, logarithmically spaced, λ-values computed in N separate simulations.

2.6 Principal Component Analysis

Principal component analysis (PCA) [78] is a simple, non-parametric technique to reduce complex data to lower dimension revealing the underlying simplified structure, if any. It is an invaluable tool in a wide variety of fields including neuroscience, data compression, computer vision/pattern recognition, data visualization, exploratory data analysis and image processing [79]. Depending on the field, it is also called discrete Karhunen-Loève transform (KLT) [80], Hotelling transform, proper orthogonal decomposition (POD), quasi-harmonic analyis [56, 81–83], or singular value decomposition [84–87].

2.6.1 Theory of Principal Component Analyis

Here, the basic ideas of PCA, together with underlying assumptions and resulting limitations, are shortly reviewed. Generally, the goal of PCA is to find the coordinate system that best re-expresses the original data, where best means maximum signal at minimal noise under a few assumptions.

Applied to an ensemble of M protein structures $\left\{\mathbf{r}^{(k)}\right\}_{k=1,\ldots,M}$, where $\mathbf{r} = (\mathbf{r}_1, \ldots, \mathbf{r}_N)$ denotes the positions of the N atoms in three-dimensional space, PCA aims at finding unit vectors $\mathbf{a}_i$ such that the projections $p_i = \mathbf{a}_i \left(\mathbf{r} - \langle\mathbf{r}\rangle\right)$ of the structural ensemble on this particular unit vector exhibits maximum variance $\sigma = (\mathbf{r} - \langle\mathbf{r}\rangle)$. These unit vectors $\mathbf{a}_i$ are then called principal modes, the projections p_i principal components. The principal components are retrieved by transforming the covariance matrix $\mathbf{C}$ of atomic fluctuations with elements $C_{ij} = \langle (r_i - \langle r_i\rangle) (r_i - \langle r_i\rangle)^T \rangle$. $\mathbf{C}$ contains two kinds of information: (i) the diagonal terms indicate interesting dynamics (noise) if the values are big (small) and (ii) the off-diagonal terms indicate high (low) redundancy if the values are high (low). PCA, consequently, diagonalizes $\mathbf{C}$,

$$\mathbf{T}^T\mathbf{C}\mathbf{T} = \mathrm{diag}\left(\lambda_1, \ldots, \lambda_{3N}\right), \qquad \text{where } \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{3N},$$

obtaining the desired principal modes and the principal components, which are just the columns of the transformation matrix $\mathbf{T}$ and $p = \mathbf{T}^T \left(\mathbf{r} - \langle\mathbf{r}\rangle\right)$ respectively.

2.6.2 Assumptions and Limitations

The underlying assumptions in this procedure are fourfold. First, PCA restricts the solution space drastically by considering only linear combinations of the original basis.Second,

it is assumed that mean and variance entirely describe the probability distribution. The only class of probability distributions that are fully described by these first two moments are Gaussian distributions. Applied to biomolecules, the modes which (almost) fulfill this assumption are called (quasi-)harmonic. The assumption, however, means that PCA breaks down for non-Gaussian distributed data. Third, it is assumed that the most interesting, principal, basis vectors are those in which the system expresses the highest variance. Fourth, the basis vectors are supposed to be orthogonal. This straightforward assumption renders the problem soluable with linear algebra decomposition techniques.

PCA has been shown to be useful in the context of theoretical biophysics. It has, for example, been shown that indeed protein dynamics is governed by a few degrees of freedom. In fact, as much as 90% of atomic displacement can be described by only 5-10% of collective coordinates, which also describe most of the functional dynamics of proteins [88, 89]. This has led to the concept of the essential subspace which has been the main target of enhanced sampling techniques [90–92] (one of which presented in Section 2.4) and models of protein dynamics [91, 93, 94]. Hence, the maximum variance assumption of PCA appears to be mostly justified.

The assumption of Gaussian-shaped densities has some justification, too. This is a consequence of the Central Limit Theorem, which states that any distribution of arithmetic means of random samples taken from any finite-variance distribution is Gaussian, if the number of samples is large enough [95]. Unfortunately, particularily the essential subspace is where the Gaussian assumption discussed above does often not hold, since the high-amplitude modes express highly anharmonic behaviour [89, 96]. PCA modes, thus, fall into two classes: essential modes with high-amplitude are normally badly described by a Gaussian, whereas non-essential modes with small amplitude are well described by a Gaussian.

2.7 Full Correlation Analysis

The goal of Full Correlation Analysis (FCA) [97] is the same as for PCA. Like PCA, it aims at finding a new set of uncorrelated basis vectors; however, FCA abandons all the assumptions of PCA except for linearity, and tries to find a coordinate system that satisfies the most general kind of non-correlation, statistical independence.

The concept of indepence roots in information theory. Two random vectors y_i and y_j are called independent if knowledge of the value of y_i does not give information of the

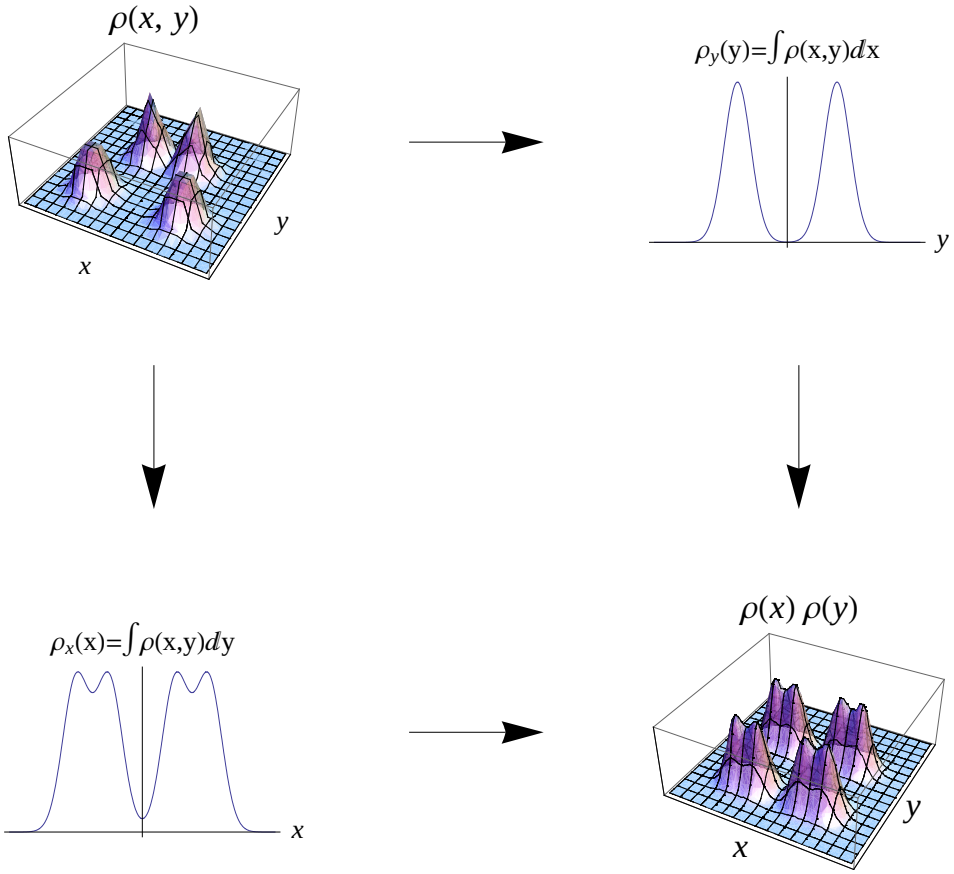

Figure 2.5: Relation between joint probability distributions (upper left), marginal distributions (upper right and lower left) and the product of marginal distributions (lower right). The entropy estimated by multiplication of the marginal PDFs is always larger than or equal to the joint entropy.

value of y_j, and vice versa. The vectors y_i and y_j are independent if and only if the joint probabilty density function (pdf) of y_i and y_j can be factorized,

$$\rho(y_i, y_2) = \rho_i(y_i)\rho_j(y_j), \tag{2.2}$$

where the marginal pdf (cf. Fig. 2.5) of y_i,

$$\rho_i(y_i) = \int \rho\left(y_i, y_j\right) dy_j. \tag{2.3}$$

A whole series of methods akin to Independent Component Analysis employing slightly different approaches to find statistically independent basis vectors has been developed [98–104]. FCA can be viewed as a specialized ICA-like tranformation method with algorithms optimized for the application to computer generated atomistic ensembles. Specifically, it finds an orthogonal coordinate transformation $\mathbf{T}$ of the M-snapshot ensemble $\mathbf{R} = \left\{\mathbf{r}^{(k)}\right\}_{k=1,\ldots,M}$ of $3N$ cartesian coordinates $\mathbf{r} = (r_1, \ldots, r_{3N})$ of the N particles such that the probability distribution of the resulting coordinates

$$\mathbf{R}' = \mathbf{TR}$$

can be factorized, i.e. like above

$$\rho(\mathbf{r}'_i, \mathbf{r}'_j) = \rho_i(\mathbf{r}'_i)\rho_j(\mathbf{r}'_j),$$

for all i and j, $i \neq j$.

FCA considers mutual information

$$I(\rho(y)) = \sum_{i=1}^{m} S(\rho_i(y_i)) - S(\rho(y)) \geq 0,$$

as a very natural measure of dependence between variables, where $\rho_i(y_i)$ are the marginal densities (Eq. 2.3) of the joint density $\rho(y)$ and S is the entropy. The entropy of a random variable can be seen as the amount of information that the observation of this variable yields. The more unpredictable (i.e. random) a variable is the larger its entropy. For a discrete random variable y, the entropy is defined as

$$S(y) = -\sum_i p(y = a_i) \ln p(y = a_i),$$

where a_i are the possible values of y. In the continuous case, with y being a random vector with probabilty density $\rho(y)$,

$$S(y) = -\int \rho(y) \ln \rho(y) dy.$$

The mutual information I employing this entropy measure is 0 if and only if all the m distributions are statistically independent. This occurs only if the joint density $\rho(y)$ is factorizable according to

$$\rho(y) = \prod_{i=1}^{m} \rho_i(y_i) \Leftrightarrow I(\rho(y)) = 0,$$

which is a generalisation of two-variable case Eq. 2.2. The mutual information takes, thus, into account the whole dependence of the variables, and not only linear dependencies like the covariance PCA and similar methods are based on. Accordingly, I is an excellent measure for the goodness of the transformation $\mathbf{R}$ operated by FCA.

Abandoning most of the simplifying assumptions of PCA and allowing a less constrained set of solutions to the problem consequently renders FCA computationally more demanding than PCA. However FCA has been demonstrated to improve significantly on the PCA for a wide range of biomolecular systems. In this thesis, it will be used as the second building block of an three-modular hierarchical method to estimating absolute configurational entropies of macromolecules (cf. Chapter 4).

Chapter 3

3 Force-induced activation of titin kinase

The first project of this thesis examines the catalytic properties of titin kinase, the catalytic domain of the muscle protein titin. The study partly presented here is the outcome of a collaboration and involved atomic force microscopy (AFM) experiments by Hermann Gaub's group, LMU München, enzymatics, conducted by Mathias Gautel's group, Imperial College, and atomistic molecular dynamics studies carried out by our group in the Theoretical and Computational Biophysics Department at the MPI for Biophysical Chemistry. The results presented here are not exclusively my work. The simulation part of the project was a close collaboration with Lars Schäfer, such that I only shouldered about half of the data production, data analysis and graphics processing required for this study. Frauke Gräter contributed the complete-TK simulations.

3.1 Titin and titin kinase

Striated muscles are composed of a series of segments interrupted by bands and disks. The basic contractile unit of muscles is the sarcomere, the segment between two neighbouring Z-disks (*Z*wischenscheibe). Sarcomeres are multiprotein complexes comprising three major filament systems, thin and thick, and connectin (cf. Fig. 3.1a). The thin filaments are predominantly composed from the globular protein *actin* coiled with nebulin filaments. The thick filament is composed of *myosin*, a motor protein, which is held in place by *titin* (connectin). When a muscle contracts, actin is pulled along myosin towards the M-band (the centre of the sarcomer). Neither actin nor myosin, which together are called the myofibril, change length during this process; only due to the increasing overlap of

actin and myosin the muscle becomes shorter (sliding filament model). While actin and myosin can thus be regarded as contracing filaments, titin is the elastic counterpart. When the muscle stretches, titin generates passive tension by straightening and partial unfolding restoring the sarcomere length and realigning the muscle filaments after stress release [105, 106] serving as an adhesion template for the contractile machinery in the muscle cells. Titin thereby keeps the filaments organized and plays a role in muscle elasticity[107].

To fulfill its duty, a single titin molecule spans half the length of the sarcomere from the Z-disks to the M-band (about $1\,\mu$m) making it the largest protein known to date with a molecular weight of approximately 3 MDa.[1] Titin is composed of a series of covalently bound domains of the immunoglobulin (Ig) and fibronectin (Fn) classes (Fig. 3.1b) which vary in sequence and number between different muscle types. Near it's C-terminus at the M-band, titin's only catalytic domain is located, the serine/threonine protein kinase domain titin kinase (TK). The crystal structure (Fig. 3.1d) shows a two-domain architecture, the smaller domain (left hand side) being β-sheet rich and the larger domain (right hand side) containing predominantly α-helices. The catalytic site is situated between the small and the large domain, where the catalytic residue, Asp127, transfers the ATP's phosphate moiety to the protein substrate.

Titin kinase activation is impeded by its C-terminal regulatory tail (red labels βR1, αR1 and αR2 in Fig. 3.1d)) by a dual autoinhibition mechanism. The $\alpha R2$ helix of the autoregulatory tail, first, blocks the ATP binding site and, second, specifically interacts with the residues crucial for catalysis. TK activation, thus, requires the autoinhibitory tail to be removed. Whereas TK activation has been shown to occur upon phosphorylation of Tyr170 and binding of calmodulin to the $\alpha R1$ helix, calmodulin or other calcium binding proteins are unable to active TK on their own [109]. Given the exceptional position of TK close to the compliant M-band, it has been speculated that TK might be activated by mechanical stress. Indeed, the M-band lattice is deformed only during active muscle contraction, and is, consequently, the ideal location for detecting the actual workload on the myofibril [110]. Frauke's previous force probe molecular dynamics simulations of the mechanical properties of TK suggested that TK activation might be possible by mechanical forces [50] and a mechanosensitive signalling complex (signalosome) was subsequently identified that interacts with an open conformation of TK and seems

[1] It is interesting to note that titin's IUPAC name, containing 189,819 letters, is sometimes stated to be the longest word in the English language [108].

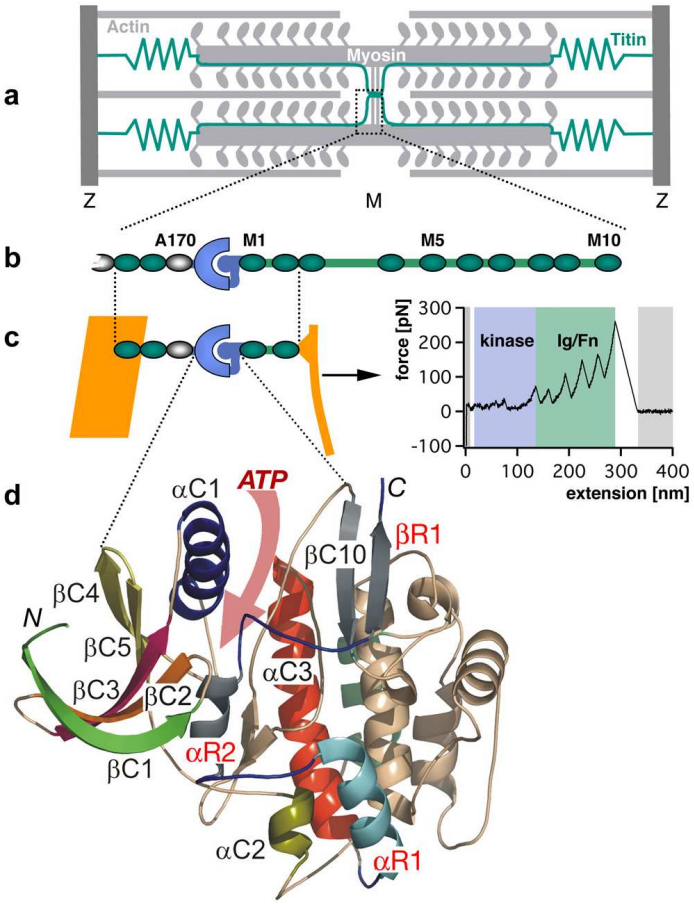

Figure 3.1: Sarcomeric location and structure of the investigated TK protein construct. (a) Schematic diagram of the sarcomere showing the transverse Z and M-bands and actin and myosin filaments, linked by the elastic titin filament. (b) Domain structure of M-band titin showing the array of structural Ig (green), Fn3 domains (white), and unique sequences (green lines) surrounding titin kinase domain (TK) (blue). (c) The titin construct contains the kinase domain surrounded by Fn3 and Ig domains. ATP binding requires relief of the C-terminal autoinhibitory tail (blue) from the active site, which can be achieved by external force. In mechanical single-molecule experiments, the construct is pulled off the gold support (yellow line) by a cantilever, resulting in a unique force spectrum when the protein is stretched and domains sequentially unfold. Analysis of the unfolding force spectrum identifies the peaks shaded in blue as kinase unfolding peaks; the five unfolding peaks shaded in green correspond to sequential Ig and Fn domain unfolding. (d) Kinase domain structure, with the ATP binding site highlighted by the pink arrow and individual secondary structure elements color-coded. Numbering is from N to C terminus, where C1 to C10 refer to catalytic core structures, and R1 to R3 (in red) refer to the regulatory tail (5). The N and C termini are marked.

to contribute to the adaption of muscle in response to mechanical strain [111]. The importance of TK in maintaining the turnover of muscle proteins is highlighted by a point mutation in the human kinase domain that causes a myopathy (muscle weakness disorder) with failure of load-dependent protein turnover [111].

Previously, no experimental proof was available that TK, and in fact any other force sensor in muscle [49], is indeed being mechanically activated rather than simply unfolded by mechanical stress. Previous in-house MD simulations, however, showed that the forces required to unfold TK are much lower than the forces required to unfold neighbouring Ig/Fn domains [50] predicting that TK, as required for a force sensor, will be completely unfolded before other titin domains. It was also shown that before complete unfolding the autoinhibition is removed, as required for mechanical TK activation. The study presented in this chapter aimed at finding a proof for these theoretical predictions. To this end, in a combined experimental and theoretical effort the connection between mechanical stress exerted on TK and its enzymatic properties was examined. Since it is most difficult to observe TK action in its natural environment firmly embedded in the contractile machinery of muscle, we conducted a study on single molecules in isolation. This allowed to examine the mechanic properties properties of TK in molecular detail using AFM experiments and MD simulations, the latter of which helped to interprete the AFM results at the atomic level and correctly predicted the residues crucial for TK catalytics. Finally, enzymatic measurements were employed to compare the single-molecule results to properties of the intact sarcomer. This chapter focuses on the MD results and the directly related AFM experiments. The full study has been published recently [112].

3.2 Simulation details

All simulations were carried out with the Gromacs simulation suite [113, 114], using the OPLS all-atom force field [64] and periodic boundary conditions. NpT ensembles were simulated with the protein and solvent coupled separately to a 300 K heat bath ($\tau_T = 0.1\,ps$) [69]. The systems were isotropically coupled to a pressure bath at 1 bar ($\tau_p = 0.1\,ps$) [69]. Application of the Lincs [31] and Settle [30] algorithms allowed for an integration time step of 2 fs. Short-range electrostatic and Lennard-Jones interactions were calculated within a cut-off of 1.0 nm, and the neighbour list was updated every 10 steps. The particle mesh Ewald (PME) method was used for the long-range electrostatic interactions [72], with a grid spacing of about 0.12 nm.

3.2.1 ATP Force Field

The atomic partial charges of ATP used along with the OPLS force field were derived from quantum chemical calculations. The charges were obtained from B3LYP/6-31+G* calculations using the CHELPG electrostatic potential fitting scheme [115]. The quantum chemical calculations were carried out with Gaussian03 [116]. All non-bonded parameters are given in the appendix A. Bonded parameters were taken from Reference [117].

3.2.2 Generation of Starting Structures

The simulations of the full-length titin kinase (TK) were set up as described in Ref. [50], with the exception that here the OPLS force field was used (see above).

The starting structures of the truncated TK were obtained from the TK crystal structure (PDB entry 1TKI, Ref. [109]) in three steps detailed below. First, 33 residues at the C-terminus comprising the αR1 and αR2 motives of the autoinhibitory tail (ai tail) were removed from the structure, since they block the ATP binding site. Second, the ATP ligand was docked into the active site. Third, the ligand-induced conformational closure of the protein structure was enforced by MD.

Docking of ATP

ATP was docked into the active site of TK using the protein kinase A (PKA, PDB entry 1q24, Ref. [118]) as a homology model. After aligning of the two protein structures, the ATP and one Mg^{2+} ion were adopted from PKA. The second Mg^{2+} ion was added using the phosphorylase kinase structure as a template (PDB entry 2phk, Ref. [119]), because it was not resolved in the PKA structure.

Energy Minimization and Equilibration of the Solvent

Prior to the free MD simulations, the systems were solvated with TIP4P water within a cubic box of 8.5 nm length. Sodium and chloride ions were added ($c = 0.15\,\mathrm{mol/l}$), and the systems were energy minimized for 1000 steps using steepest descent. The solvent was then equilibrated for 500 ps with positional restraints on the protein heavy atoms (force constant $1000\,\mathrm{kJ\ mol^{-1}\,nm^{-2}}$).

Closure of the Active Site

The ligand-induced conformational closure of the protein structure was enforced by means of essential dynamics MD (see Sec. 2.4). To this end, the closed conformation of the PKA (1q24) was used as the target structure. First, the two protein structures (TK and PKA) were aligned. Then, the N-terminal β-hairpin (residues 13-21 in TK) was selected, and a principal component analysis (PCA) was carried out, yielding one eigenvector that describes the closing motion. Essential dynamics sampling was then performed, during which motion along the eigenvector towards the target structure was enforced (radcon option in EDsampling module [120] of Gromacs). To allow the protein to relax along the enforced closing motion, the maximal step size along the eigenvector was restricted to a maximum of 0.05 pm/step, leading to a closed structure within about 1 ns. Subsequently, the closed structure was simulated for 1 ns with positional restraints on the Cα atoms (force constant 1000 kJ $\text{mol}^{-1}\,\text{nm}^{-2}$). Finally, a 1 ns free MD simulation was carried out, during which no opening motion was observed in the presence of ATP. As a control, the closed structure was simulated also in the absence of ATP. As expected, significant re-opening motion was observed. The starting conformation for the force-probe MD simulations were taken from the final free MD simulations after 1 ns.

In the resulting structure with ATP bound, a salt bridge between Lys36 and the α-phosphate group of ATP, and interaction between Met34 and the adenine moiety, respectively, were formed. For the latter, two types of interactions with ATP are seen, a hydrogen bond between the sulphur atom and the NH2 group of adenine, and, alternatively, a stacked conformation of the S-CH_3 group and the adenine 6-ring. Additional hydrogen bonds were formed between the β-phosphate of ATP and the N-H backbone of Glu17, as well as between the hydroxyl groups of the ribose and the carbonyl backbone of Arg15.

3.2.3 Force-Probe MD Simulations

To mimic the AFM experiments, force-probe MD simulations (FPMD) were carried out as described in methods. The average force at the two springs was monitored during the simulations, yielding the force profiles shown in Fig. 3.3B.

Prior to the FPMD simulations, the protein was aligned along the pulling direction (z-axis), and the simulation box was extended along the z-axis to about 20 nm, allowing for the accommodation of an elongated conformation. Further, water, sodium, and chloride ions ($c = 0.15\,\text{mol/l}$) were added, and the system was energy minimized, followed

by equilibration for 200 ps with positional restraints on the protein heavy atoms (force constant 1000 kJ $\mathrm{mol}^{-1}\,\mathrm{nm}^{-2}$).

To generate statistically independent unfolding trajectories (five trajectories with and without ATP, respectively), a partially unfolded conformation was selected after 19.2 ns and 20.2 ns of FPMD simulation time with and without ATP, respectively. These conformations were chosen, because they correspond to minima in the force profiles. The nitrogen atom of the N-terminus and the carbon atom of the C-terminus, respectively, were kept fixed with a positional restraint (force constant 1000 kJ $\mathrm{mol}^{-1}\mathrm{nm}^{-2}$), and the systems were equilibrated for 1 ns with free MD. From this trajectory, five equidistant frames ($\Delta t = 200$ ps) were chosen as starting structures for the additional force-probe simulations.

Simulations were interrupted before any of the springs had crossed the box boundary. Here, care was taken that, due to the applied periodic boundary conditions, the pulled termini did not interact with each other. At this point, fully unfolded residues at both termini were removed, and new termini, water, and ions were added. The FPMD simulations were then continued after equilibration of the solvent with positional restraints on the protein heavy atoms (force constant 1000 kJ $\mathrm{mol}^{-1}\mathrm{nm}^{-2}$) for 200 ps.

The simulated systems comprised about 180,000 atoms. The total simulation time was about 500 ns.

3.2.4 Contour Length Histograms

The plotted contour lengths aim at quantitative comparison between the simulations and the experiments. Such comparison is complicated by the fact that (a) different pulling speeds are used, and (b) in the simulations, a stiffer spring has to be used then in the AFM experiments. To account for these differences, we derived contour length plots from the force profiles obtained from the MD simulations in three steps. First, prominent force peaks were selected along the force profiles. Here, only those force peaks j were included whose height $F_{\mathrm{max},j}$ exceeded a certain threshold. This threshold was defined as

$$F_{\mathrm{max},j} \geq F_{\mathrm{max},j-1} - k_{\mathrm{AFM}} \left(\Delta z_{\mathrm{spring},j} - \Delta z_{\mathrm{spring},j-1} \right),$$

where k_{AFM} is the effective spring constant of the AFM cantilever and the attached linkers, and $\Delta z_{\mathrm{spring}}$ the distance between the two springs attached to the C- and N-terminus, respectively, and $j-1$ denotes the force peak preceding peak j. Here, k_{AFM} was estimated

as 10 pN/nm, which is an upper limit. This selection procedure mimics the effect of the soft AFM cantilever, which, in contrast to the force probe simulations, is insensitive to minor force peaks that immediately follow a larger peak. Second, the number of unfolded residues $N_{\mathrm{unf},j}$ and the length of the folded core $z_{\mathrm{folded},j}$ were determined for the conformations corresponding to each of these peaks. The contour length $l_{\mathrm{c},j}$ was then calculated as $l_{\mathrm{c},j} = 0.365 N_{\mathrm{unf},j}\,\mathrm{nm} + z_{\mathrm{folded}}$. Third, a Gaussian function was assigned to each of the peaks, weighted by its height $F_{\mathrm{max},j}$:

$$G_j(l_c) = F_{\mathrm{max},j} \exp\left[-\frac{(l_c - l_{c,j})^2}{2\sigma^2}\right],$$

with a width $\sigma = 1\,\mathrm{nm}$. The final contour length plot is obtained from the sum of all these Gaussian functions. As an exception, this procedure was not applied to the force peak marked with a plus sign in Fig. 3.3B (upper row). Structural analysis shows that peak arises from transient rupture and reformation of the interactions between Lys-36/Met-34 and ATP, similar to the subsequent ATP peak 2*. From the considerable scatter of heights seen for peak 2*, we expect that the peak in question is exceptionally pronounced and, therefore, that the described shadowing effect will not occur in most of the cases. No other peak showed a similar effect.

To allow comparison with contour histograms derived from AFM experiments, one has to consider that the AFM measurements were conducted with a full TK that additionally includes a linker of 23 residues. No structural information was available for this 23-residue linker between A170 and the catalytic core. In the homologous twitchin kinase, this sequence wraps around the ATP-binding lobe of the kinase, but makes no contacts with the autoregulatory domain. We therefore expect this part to affect the relative positions of the observed unfolding peaks, but not the sequence of local unfolding events at the C-terminus.

In the MD simulations, either a full-length TK or a truncated TK, i.e., with a part of the autoinhibitory tail removed, were used. As a reference point for comparison, peak 6 was chosen, which indicates unfolding of Ig-domain "handles" at a fully unfolded TK ($l_c = 144.4\,\mathrm{nm}$). The 321 residues of the full TK, or the 289 residues of the truncated TK would all be unfolded at this point.

3.3 Results

Experimental reference

The AFM measurements were carried out with a TK construct comprising the protein kinase domain plus neighbouring Ig/Fn domains as sketched in Fig. 3.1 and shown in more detail in Fig. 2.2. The resulting forces were recorded with pico-Newton accuracy and showed a typical saw-tooth pattern as explained in Sec. 2.3.

Typically, a series of five initial low-force peaks (below 50 pN) was followed by up to five distinct saw-tooth-shaped high-force peaks that correlated exactly with the number and contour lengths of the flanking Ig/Fn-domains. These low-force peaks, occuring before Ig/Fn unfolding, stem from unfolding events within the kinase domain. The forces required for complete unfolding of the TK domain does not exceed 50 pN at 23°C and a pulling speed of 1 μm/s, as predicted from previous MD simulations [50]. From the fact that the mechanically more stable Ig/Fn domains always unfold after the TK domain it is obvious that the force acts on all domains in series, such that the kinase domain is therefore completely stretched before the first Ig/Fn domain unfolds.

Fig. 3.2 shows experimental force profiles obtained with and without ATP. As can be seen, ATP binding entails an additional force peak (see Fig. 3.2 C, peak 2*). This peak depends on ATP concentration as well as the time span between binding site opening and the moment the ATP barrier (peak 2*) is probed. It is, thus, clearly due to ATP binding[112].

Molecular Mechanism of TK Activation by Force

We used force–probe MD simulations [75, 121] to characterize the force-induced unfolding of TK at the atomic level and to correlate the structural states with the energy barriers observed by the single-molecule force spectroscopy experiments. Force–probe molecular dynamics simulations [75, 121] used the TK x-ray structure [Protein Data Bank entry 1TKI [109]] as the starting structure, with the autoinhibitory tail partly removed (see Sec. 3.2.2). Two sets of simulations (five each) were carried out for this truncated TK: one set with an empty binding pocket, and one set with an ATP molecule and magnesium ions inserted into the (closed) binding pocket. As a control, the autoinhibited complete TK was also subjected to force–probe MD simulations (see Sec. 3.2.3). As in the experiment, the two force profiles obtained from the simulations of the truncated TK (Fig. 3.3B Top

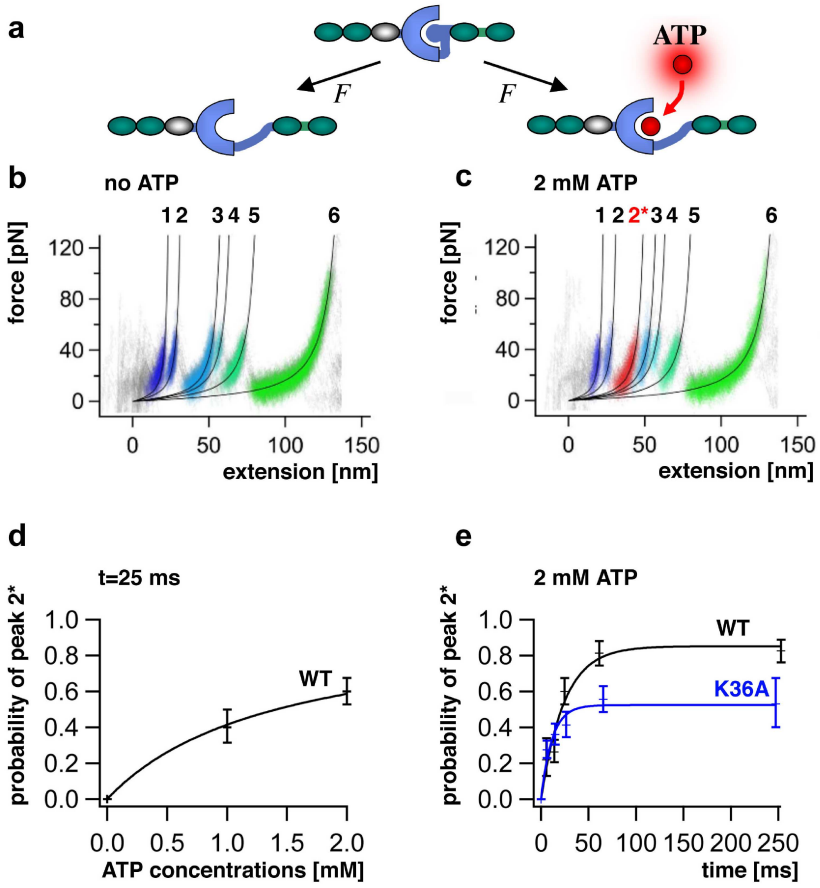

Figure 3.2: Unfolding profile of TK. (A) External force can open the ATP binding site of TK by unfolding of the autoinhibitory domain (blue ball). (B) Superimposed traces of 66 single-molecule unfolding events in TK show a fixed sequence of unfolding events, numbered 1–5. (C) Mechanically induced ATP binding leads to a distinctly altered force profile with the appearance of an extra force peak, 2*, absent in unfolding events in absence of ATP (44 traces).

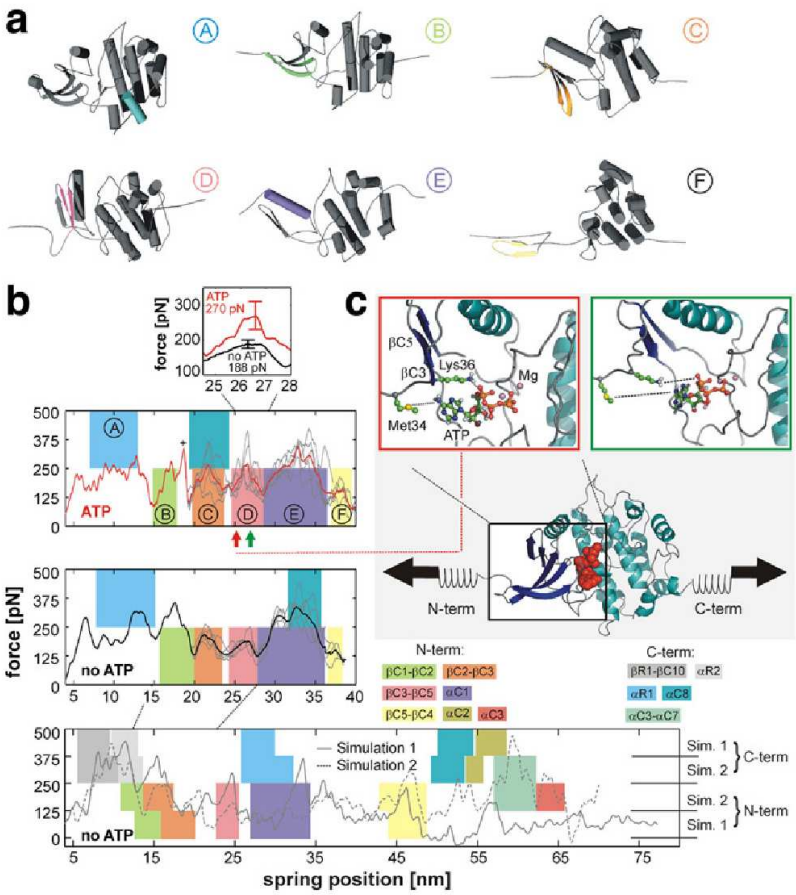

Figure 3.3: Molecular dynamics (MD) simulations of the force-induced unfolding of titin kinase (TK). (a) Representative unfolding intermediates with the unfolding secondary structure elements colored according to the scheme in Fig. 1d; the beta-strands unfold pairwise, and colors refer to the respective N-terminal strand of each pair.(b) Unfolding forces of truncated TK with ATP (upper row), without ATP (middle row), and of the complete TK (lower row). For the complete TK, two independent 90 ns simulations were carried out (solid and dashed lines). Starting from a partially unfolded structure at ˜19 nm, five 26 ns trajectories (thin grey lines) were averaged for both sets of simulations (thick lines in upper and middle row). Color shaded areas indicate main unfolding events, which correspond to the colors used in (a) and in Fig. 1d. An additional force peak in the presence of ATP is predicted (plus-sign and pink-shaded area in upper row). This force peak (inset) is higher for bound ATP (270 pN) than for an empty binding pocket (188 pN). Due to the necessarily much faster pulling rates of 0.8 m/s used for the simulations, larger unfolding forces are seen, which can be related to the experimental loading rates [50] (c) In the force-probe MD simulations, harmonic springs were attached to the protein and retracted with constant velocity (lower schematic, ATP shown as red spheres). (c, insets) Representative structures shortly before (left) and after (right) the ATP force peak. ATP and the two key residues methionine-34 and lysine-36 are shown in ball-and-stick representation, and the rupture of molecular interactions is indicated by dotted lines.

and Middle) are largely similar. A notable exception is the more pronounced force peak seen in the presence of ATP (see Fig. 3.3B Inset) at the position of the measured force peak 2*.

To allow direct comparison of the unfolding pathways between experiment and simulation, we transformed the force extension traces of Fig. 3.2 into barrier position histograms [109] and derived the same from our simulations (see Sec. 3.2.4). Figure 3.4 compares the contour lengths from experiment and MD simulation. The simulations allowed assigning the ATP peak (peak 2*) to the anchor point Lys36. Thus, 289-36=253 residues unfold between peaks 2* and 6, which corresponds to a contour length increment of 253x0.365 nm=92 nm. Including the length of the folded core of 5.5 nm, this estimate yields an l_c for peak 2* of $144.4\,\mathrm{nm} - 92\,\mathrm{nm} + 5.5\,\mathrm{nm} = 57.9\,\mathrm{nm}$ from the simulations, in good agreement with the value of 51.6 nm from the AFM.

The above procedure rests on two assumptions. First, we assume that before peak 2*, all N-terminal residues before Lys36 unfold. In addition, by simulating a truncated TK, we assume that unfolding of the omitted part of the autoinhibitory tail is uncoupled from the other unfolding events and precedes peak 2*. Both assumptions are corroborated by the simulations of the complete TK, where such an unfolding behaviour was indeed observed (see Fig. 3.3).

The simulations of the complete TK also show that peak 2 can be assigned to the unfolding of the autoinhibitory tail (Fig. 3.3). Hence, the first part of the contour length profile shown in the inset of Fig. 4 is based on these simulations. Here, Tyr7 was chosen as an anchor point, a residue that lies at the beginning of αR1. Following the same argumentation as outlined above, this choice yields a barrier position for peak 2 of 35.5 nm, in good agreement to the 32.3 nm obtained from the AFM data. We note that, in contrast to peak 2*, where the anchor point can be easily assigned to the Met34/Lys36 motive, such a detailed assignment is more challenging for peak 2. Here, only a range of αR1-residues between Glu5 and Asp12 can be specified, thus leading to an uncertainty in the position of peak 2 of about 1.5 nm.

Figure 3.5 compares the barrier positions obtained from the AFM experiments and the two independent MD simulations of the complete TK in the absence of ATP (Fig. 3.3B, lower row). The MD simulations yield positions of peaks 2, 3, 4, and 5 at 35.5 nm, 63.3nm, 80.1 nm, and 86.8 nm, respectively. The corresponding peak positions obtained from the AFM experiments are 32.3 nm, 61.3 nm, 68.6 nm, and 86.6 nm, respectively. For the force peaks 2, 3, and 5, the positions obtained from the simulations agree well

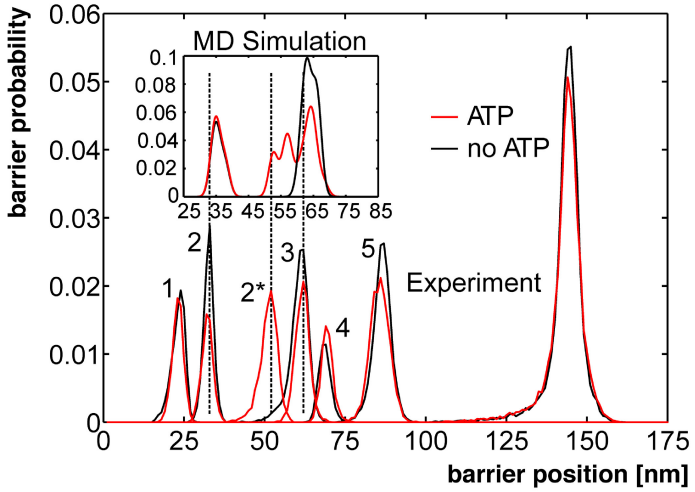

Figure 3.4: Contour length histograms obtained from single-molecule force spectroscopy experiments (transformation with QM-WLC and P=0.8 nm) and from MD simulations (Inset). The folded kinase construct has a length of 25 nm. The peak positions with (red) and without (black) ATP are similar in both histograms (dashed lines), except for one additional peak in the presence of ATP (red peak at ≈51.6 nm). The experimentally determined contour length increments are, in the absence of ATP, 9.1, 28.6, 7.3, 18.0, 57.9 nm; and, in the presence of ATP, 9.1, 19.4, 10.1, 7.5, 16.4, 58.3 nm—with an estimated error of ±2%. The position of the initial peak (24 nm) reflects the mean length of the TK construct with completely folded domains.

with those obtained from the experiments, taking into account that the former are based on only two independent simulations. For peak 4, the difference between the simulations and the experiments is somewhat larger. We speculate that this difference results from the underlying free energy landscape that, at such partially unfolded conformations, is shallow and exhibits many barriers of comparable height along various unfolding directions, such that the pathways followed in the individual trajectories are different. Such heterogeneity of unfolding pathways is indeed seen from the color-coded unfolding events in Fig. 3.3B, lower row. A much larger number of trajectories would be required to thoroughly characterize this region of the energy landscape, which, however, is not the focus of this work.

In summary, the two histograms agree well both in the presence and absence of ATP (Fig. 3.4 (dashed lines) and Fig. 3.5), allowing the conclusion that the main unfolding

events are correctly described by the simulations.

Next, we investigated which molecular interactions determine the observed force peaks. For the ATP peak 2*, two strong interactions are seen, a salt bridge from lysine-36 to the α-phosphate group of ATP, and a contact between methionine-34 and the adenine moiety of ATP (Fig. 3.3C). Both interactions break irreversibly upon βC3-βC4 rupture, giving rise to the significantly larger force peak of 270 ± 39 pN in the simulations with bound ATP as compared with 188 ± 13 pN without ATP (Fig. 3B Inset). Notably, in the AFM experiment, the contour length of 51.6 nm for the ATP peak position (Fig.3.4, peak 2*) also points to a residue close to lysine-36. Moreover, subsequent kinetic experiments of a mutation of lysine-36 to alanine (K36A) resulted in a dramatic decrease of TK affinity towards ATP, corroborating the central role of lysine-36 seen in the simulations. An additional peak is seen at 18 nm for the simulation with ATP present (plus sign in Fig. 3.3B Top). Here, a force-induced deformation of the N-terminal domain triggers the transient rupture and reformation of the methionine-34–ATP and lysine-36–ATP interactions.

Closer structural analysis of our simulations suggests the following sequence of events (colors in Fig. 3.3A and B, and Fig. 3.4). Peak 1 (Fig. 3.4) is caused by unfolding of the 23-residue linker at the N terminus of TK, which is not present in the simulations (see 3.2.2). At peak 2, the autoinhibitory tail is unfolded and removed, rendering the ATP binding site accessible (region shaded in gray in Fig. 3.3B Bottom). Subsequently, N-terminal β-sheets βC1-βC2 and βC2-βC3 rupture (regions B and C). For these events, no force peak is seen in the experiment, because it would fall into the lag time after force peak 2. Peak 2* described above is dominated by interactions of ATP with the binding pocket. The truncated construct necessarily lacks part of the autoinhibitory tail stabilizing the adjacent C-terminal α-helix αR1 in the full-length TK. Accordingly, αR1 unfolds first in the truncated kinase (Fig. 3.3B Top and Middle, region A) but after βR1 and αR2 in the autoinhibited kinase (Fig. 3.3B Bottom). Hence, and in agreement with the complete TK unfolding simulations (Fig. 3.3B Bottom), peaks 3 and 4 are assigned to unfolding of αC1 and βC4-βC5, respectively (regions D and E). Finally, peak 5 arises from the combined effect of αC2 and αC8 rupture (Fig. 3.3B Bottom). At peak 6, the complete TK is unfolded and stretched. Taking the diameter of the folded TK into account (5.5 nm), the contour length increment to peak 1 (121 nm) corresponds to $(5.5 \pm 121 \pm 2)$nm/0.365 nm$= 346 \pm 5$ residues, in agreement with the 344 aa of TK including its N-terminal linker.

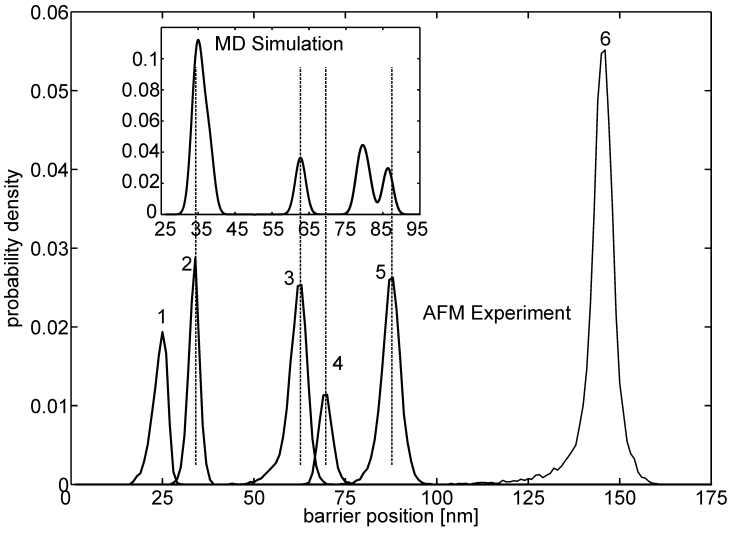

Figure 3.5: Barrier positions from AFM experiments and MD simulations in the absence of ATP. The latter were obtained from the simulations of the complete TK (321 residues) shown in Fig. 3.3B (lower row). The positions of the force peaks obtained from the AFM experiments are indicated by the dashed lines.

3.4 Discussion

Experimental and simulation results together show that TK can be activated by mechanical stress and, thus, act as a molecular force sensor in the sarcomers of striated muscles. In the force probe experiments, the activation forces are within the physiological range and, importantly, lower than the ones unfolding the surrounding titin domains. Small force imbalances of four to eight myosin motor domains, $\approx 3\%$ of the myosin molecules between adjacent thick filaments, could thus translate into a physiologically significant signal by activation of the TK mechanosensor. The molecular dynamics simulations carried out for this study provided valuable insight into the molecular basis underlying this force sensor. Using contour plots, we were able to make direct contact with the experimental force profiles and showed that, despite the neccessarily much higher forces required in the force-probe simulations, the main events were correctly predicted in the simulation and that the 2* peak seen in the AFM measurements indeed corresponded to TK interactions with ATP. Notably, the central residue to this mechanism, Lys-36, which was identified in the simulations, indeed turned out to be crucial for catalysis; the K36A mutation subsequently examined was almost devoid of any catalytic activity. The biological implications of the overall results are manifold and include enhanced understanding of the consequences of eccentric exercise, and speculations that TK mechano-activation might constitute a paradigm also for other members of the family of cytoskelatal autoregulated protein kinases [112].

Chapter 4

4 Estimating configurational entropies: The minimally coupled subspace approach

The work presented here and in the following two chapters was done in close collaboration with Oliver Lange. None of the work presented in this and the following chapters is exclusively my work. FCA was developed by Oliver, and its basic concepts are described in the methods chapter. Oliver also wrote the density estimation program *g_entropy* and coordinated the project as supervisor-in-charge. I selected the test systems, performed the simulations, analyzed the data, created the figures and established the third building block of MCSA, the mutual information expansions introduced later in this chapter and, in more detail in Chapter 6. This chapter provides an overview on the method and first applications of all of the modules on biological macromolecules. The density estimation module and the mutual information expansions are derived and presented in more detail in the following two chapters.

4.1 Introduction

Entropies are key quantities in physics, chemistry, and biology. While free energy changes govern the direction of all chemical processes including reaction equilibria, entropy changes are the underlying driving forces of ligand binding, protein folding or other phenomena driven by hydrophobic forces. Atomistic simulations, e.g., molecular dynamics, in principle provide all the information needed for calculating both, free energies and entropies. Yet calculating entropies from atomistic ensembles $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ of n configurations $\mathbf{x}_i \in \mathbb{R}^{3N}$

of a macromolecule of N atoms remains notoriously difficult. In this chapter, we develop and apply a method for calculating configurational entropies

$$S_c \sim -\int \rho\left(\mathbf{x}\right) \ln \rho\left(\mathbf{x}\right) d\mathbf{x}, \tag{4.1}$$

where $\rho(\mathbf{x})$ denotes the configurational probability density $\rho(\mathbf{x}) = \exp(-\beta V(\mathbf{x}))/Z_c$ in the $3N$ dimensional configurational space governed by the potential energy $V\left(\mathbf{x}\right)$. The fact that N is usually in the order of several hundreds or thousands renders the evaluation of this integral quite challenging despite a number of attempts [53, 54, 122, 123], which broadly fall into three classes, (i) special-purpose perturbation type approaches, also known as thermodynamic integration [124], (ii) step-by-step reconstruction methods, in particular the scanning procedures introduced by Meirovitch [125, 126], and (iii) direct approaches which analyse information readily available in standard equilibrium simulation trajectories [56, 57, 127].

While perturbation approaches provide relatively accurate free energy differences also for larger systems, accurate entropies are obtained only for smaller molecules. The main obstacle, which aggravates with system size, is the sampling problem, which severely limits the accuracy, in particular for explicit solvent models [122]. On the one hand, for free energy differences, perturbation type approaches exploit cancellation of terms in the difference of the partition functions of the two states. Thus, it suffices to sample those degrees of freedom which are affected by the perturbation which are usually few (e.g. regions around binding sites). If the two states differ substantially, however, sampling suffers from slow convergence.

For entropy differences, in contrast, the full Hamiltonian contributes to the result, such that here complete sampling of the full phase space is indeed required [124]. Thus, perturbation or scanning approaches do not share the same fundamental advantage over a direct evaluation of the partition sum to obtain entropies as they do for free energy differences.

Direct approaches, in principle, have additional advantages. First, such methods are, by construction, independent from finding suitable perturbation pathways between the states of interest and/or analytical tractable reference states. Furthermore, to gain a detailed understanding of the molecular mechanisms, one would like to separate various contributions to the entropy (changes), e.g., side-chain, backbone, or ligands. With direct methods, all these analyses could be carried out on the same computational ensembles.

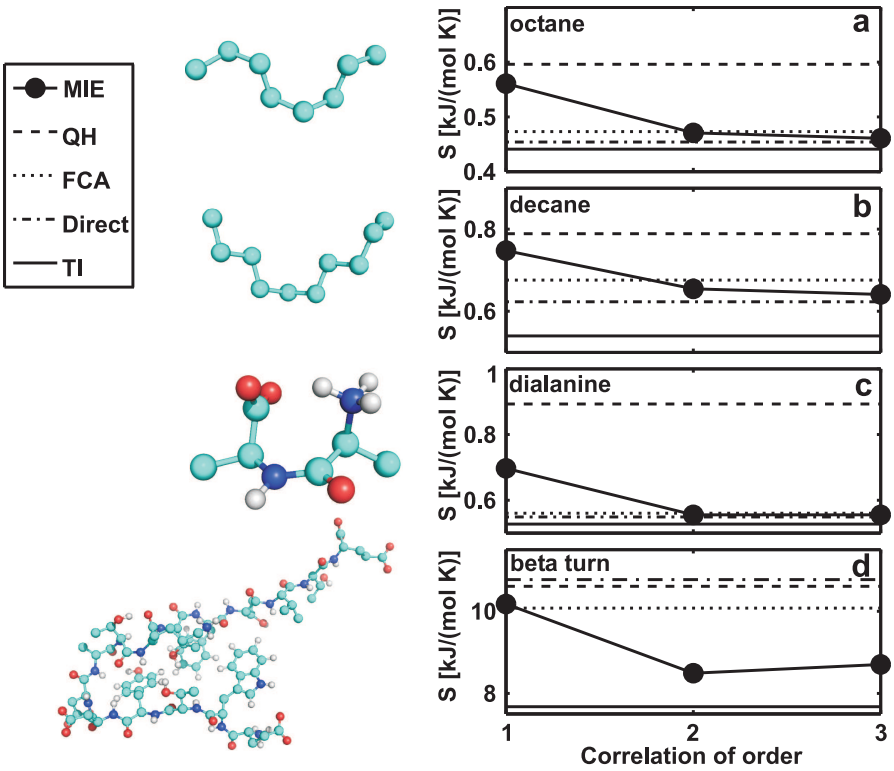

Figure 4.1: Entropy estimates for two selected alkane systems, for dialanine and for the last $\beta-$turn of Protein G. Thermodynamic Integration values (*TI*; horizontal solid lines), density estimates without over the whole configurational space (*Direct*; dashed-dotted lines), Full Correlation Analysis (*FCA*) with subsequent clustering and density estimation (dotted lines), quasi-harmonic estimate (*QH*; dashed lines) and Mutual Information Expansion values (*MIE*; black circles) were obtained as described in the text. *MIE* values include correlation orders 1, 2, and 3, respectively.

The most widely used direct method is the quasi-harmonic approximation (QH) [56], which renders the integral Eq. (4.1) separable. Accordingly, the (quasi-) harmonic entropy $S_{c,\mathrm{harm}} = \sum_i^{3N} S_c(i)$ is given as a sum of entropies $S_c(i)$ of the individual (quasi-) harmonic modes, which is equivalent to approximating the configurational density $\rho(\mathbf{x})$ by a multi-variate Gaussian function, $\rho(\mathbf{x}) = (2\pi)^{-3N/2} \det \mathbf{A} \exp\left(-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}\right)$ with $\mathbf{A^{-1}} = \mathbf{C}$ derived from the covariance matrix [57, 127] $\mathbf{C} = (n-1)^{-1} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}$. Using the partition sum of the quantum mechanical oscillator for the entropies $S_c(i)$ of individual modes, this estimate has been shown to provide a rigorous upper limit of the true entropy [57]. However, for macromolecules the true configurational entropy is typically considerably smaller due to coupling of motional modes and anharmonicity, i.e., multi-modality. Although this problem has already been pointed out quite early [56, 57], little is known about how large the overestimation actually is. Indeed, model calculations on systems as the ideal gas [128], Lennard-Jones fluids [128], or small linear alkane molecules [129], however, reveal drastic differences between the quasi-harmonic entropy estimates and the actual entropy. For complex macromolecules, this effect must be expected to be even more pronounced.

4.2 The Minimally Coupled Subspace Approach

Figure 4.1 shows that indeed for various small test systems (alkanes, dialanine and a complete 14-residue β-turn) the quasi-harmonic approximation severely overestimates the reference entropy. The reference values were obtained by thermodynamic integration (TI) gradually perturbing the systems towards an analytically tractable reference state consisting of non-interacting particles in harmonic wells, as described in methods. Similar protocols have been used elsewhere [130].

Non-parametric density estimation

Here we develop a direct method consisting of three building blocks that yields improved upper bounds to the quasi-harmonic approximation. To capture sufficient details of the configurational density we resort to non-parametric density estimation as the first building block. The configurational part of the entropy in a d-dimensional space is estimated from

n configurations according to

$$S_c = \frac{k_B}{n} \sum_{i=1}^{n} \ln \frac{n Z_d(\mathbf{x}_i) r_{i,k}^d}{k(\mathbf{x}_i, r_{i,k})},$$

where $k(\mathbf{x}_i, r_{i,k}) = \langle K(\mathbf{x}_i, (\mathbf{x}_i - \mathbf{x})/r_{i,k}) \rangle_{\mathbf{x}}$ denotes the ensemble average of a locally adapted kernel function K, whose anisotropy and scaling $r_{i,k}$ depends on the local density at point $\mathbf{x}_i$, and whose L_1-measure is denoted by $Z_d(\mathbf{x}_i)$. This formula simplifies to the well-known k-nearest neighbour entropy (k-NN) by fixing the kernel function to an (isotropic) sphere whose radius $r_{i,k}$ is chosen such that exactly k configurations are within the sphere around configuration $\mathbf{x}_i$. In this case, Z_d is the volume of the d-dimensional unit sphere. Such k-NN estimators have been shown to be applicable for up to 10 dimensional configurational space [131]. In contrast, as can be seen in Fig. 4.1(a–c), locally adapted kernels (c.f. next chaper) yield accurate results even for the 45-dimensional configurational space of dialanine. For the more than 500-dimensional configurational space of the 14-residue β-turn, however, the 'curse of dimensionality' [132] renders it impossible to improve on the quasi-harmonic approximation with direct density estimation alone.

Full Correlation Analysis

Thus, as the second building block of our method, we apply an entropy invariant transformation $\mathbf{T}$ such that the usually highly coupled degrees of freedom separate into optimally uncoupled subspaces each of which being sufficiently small to render non-parametric density estimation applicable. As the most straightforward class of entropy invariant transformations, we consider here linear orthonormal transformations of the form $\mathbf{y} = \mathbf{T}\mathbf{x}$, with $\det \mathbf{T} = 1$. More general transformations are explored elsewhere [133]. Here we apply Full Correlation Analysis (FCA) [97] which minimizes mutual information by considering

$$H[\mathbf{T}] = -\frac{k_B}{\ell} \sum_{i=1}^{3N} \int \rho_i^{(1)}(y_i) \ln \rho_i^{(1)}(y_i),$$

where y_i denotes the components of $\mathbf{y} = \mathbf{T}\mathbf{x}$ and $\rho_i^{(1)}(y_i) = \ell^{3N-1} \int \rho(\mathbf{y}) dy_{j \neq i}$ the 1-dimensional marginal density along y_i. This procedure minimizes non-linear correlations of second and higher order [97] and therefore improves on principal component analysis

(PCA) which only considers linear correlations of second order. For complex macromolecules, even for the optimal linear transformation $\mathbf{T}$, considerable correlations between several degrees of freedom will remain and cannot be neglected. To address this issue, the FCA modes are subsequently clustered according to the generalized correlation coefficient

$$r_{\mathrm{MI},ij} = \left(1 - \exp\left[-2I_{i,j}^{(2)}\right]\right)^{1/2},$$

with the mutual information

$$\begin{aligned} I_{i,j}^{(2)} &= H_i^{(1)}[\mathbf{T}] + H_j^{(1)}[\mathbf{T}] - H_{i,j}^{(2)}[\mathbf{T}] \\ &= -\frac{k_\mathrm{B}}{\ell} \int \rho_{i,j}^{(2)}(y_i, y_j) \ln \frac{\rho_{i,j}^{(2)}(y_i, y_j)}{\rho_i^{(1)}(y_i)\, \rho_j^{(1)}(y_j)} \end{aligned}$$

between components y_i and y_j. This is achieved by assigning mode indices j to m clusters C_s such that all modes with correlation coefficients larger than a certain threshold θ are assigned to the same cluster. This disjoint clustering defines an approximate factorization $\rho(\mathbf{y}) \approx \prod_{s=1}^{m} \rho_s^{(d_s)}\left(\bigotimes_{j\in C_s} y_j\right)$, where $\rho_s^{(d_s)}$ denotes the generalised d_s-dimensional marginal density along $\bigotimes_{j\in C_s} y_j$. This factorization is approximate in the sense that for the entropy

$$S[\rho(\mathbf{y})] = \sum_{s=1}^{m} S\left[\rho_s^{(d_s)}\left(\bigotimes_{j\in C_s} y_j\right)\right] + S_\mathrm{res}\left[\{C_s\}_{s=1,\ldots,m}\right]$$

the residual entropy $S_\mathrm{res}\left[\{C_s\}_{s=1,\ldots,m}\right]$ is small.

Such approximate factorization, of course, neglects the inter-cluster pairwise correlations as well as higher-order correlations. We therefore have to assume that our threshold criterion also serves to eliminate higher-order correlations between clusters. This assumption is supported by the observation that for the alkanes and dialanine, where $\theta = 0.025$, $S_\mathrm{dir} \approx S_\mathrm{FCA}$ (where S_dir is a direct density estimate over the whole configurational space without further subspace clustering; cf. Fig. 4.1a–c), i.e. that our factorization yields accurate entropies and S_res is indeed small.

Mutual Information Expansion (MIE)

However, for larger molecules, the necessarily small threshold typically results in at least one cluster that is too large (e.g. for the β-turn $d_1 = 108$) for a sufficiently accurate density estimate. Accordingly, while our factorization does improve the entropy estimate (c.f. Fig. 4.1d) still S_{residual} cannot be neglected anymore. The third building block of our method addresses this issue by subdividing each oversized cluster into h_s disjoint subclusters $D_i^{(s)}$ of sizes $d_1^s, \ldots, d_{h_s}^s < 15$, $C_s = \bigcup_{i=1}^{h_s} D_i^{(s)}$, irrespective of the necessarily remaining strong correlations between these . The residual entropy contributions to the configurational entropy

$$S\left[\rho\left(\mathbf{y}\right)\right] = \sum_{s=1}^{m}\sum_{a=1}^{h_s} S\left[\rho_s^{(d_s)}\left(\bigotimes_{j\in D_a^{(s)}} y_j\right)\right] + \sum_{s=1}^{m} S_{\text{res}}\left[\left\{D_a^{(s)}\right\}_{a=1,\ldots,h_s}\right] + S_{\text{res}}\left[\{C_s\}_{s=1,\ldots,m}\right]$$

will be drastically increased due to non-neglegible intra-cluster contributions $S_{\text{res}}\left[\left\{D_a^{(s)}\right\}_{a=1,\ldots,h_s}\right]$ from all subdivided clusters C_s, where we have omitted the argument ρ in the rightmost two terms for brevity. We here propose to compute each $S_{\text{res}}\left[\left\{D_a^{(s)}\right\}_{a=1,\ldots,h_s}\right]$ via the mutual information expansion (MIE) as

$$S_{\text{res}}\left[\left\{D_a^{(s)}\right\}_{a=1,\ldots,h_s}\right] = -\sum_{a<b} I_2^{(d_a^s+d_b^s)}\left[\rho_a,\rho_b\right] + \sum_{a<b<c} I_3^{(d_a^s+d_b^s+d_c^s)}\left[\rho_a,\rho_b,\rho_c\right] - \ldots (-1)^{h_s+1} I_{h_s}[\rho_a,\ldots,\rho_{h_s}],$$

where $\rho_a \equiv \rho^{(d_a)}\left(\bigotimes_{j\in D_a^{(s)}} y_j\right)$. Expanding the mutual information terms

$$I_k^{(\sum_1^k d_a)}\left[\rho_1,\ldots,\rho_{h_s}\right] = \sum_{a=1}^{k}(-1)^{a+1}\sum_{i_1<\cdots<i_a} S[\rho_{i_1},\ldots,\rho_{i_a}],$$

up to second or third order, respectively, with the right-hand sum running over all possible permutations $\{i_1,\ldots,i_a\} \in \{1,\ldots,k\}$, has proven sufficiently accurate in liquid state

theory [134, 135]. Indeed, for the β-turn, inclusion of the remaining correlations via this expansion improved the entropy estimate to near agreement with the TI reference (Fig. 4.1d).

We note that previous attempts to apply MIE to macro-molecular systems have failed [136, 137] due to combinatorial explosion and slow convergence. Here, this problem is solved by clustering into < 15-dimensional subspaces which minimizes these correlations and delays the onset of combinatorial explosion. This criterion also guarantees that even for the third-order MIE no direct density estimates beyond the critical dimensionality $d_s = 40$ are required.

Application to Biomolecules

Together, these three building blocks should enable one to calculate configurational entropies even for larger biomolecules. As an example of biological relevance, the 146-residue globular protein calmodulin was considered for which scanning or perturbation methods do not converge. Calmodulin (CaM) is a calcium regulated protein (Fig. 4.2) which binds up to four calcium ions (yellow) under physiological conditions. Only the calcium bound structure binds many different peptide binding partners with high affinity. As no direct interaction of calcium with the peptide binding partnes occurs, lowering the configurational entropy of the free (but calcium bound) calmodulin has been suggested as a possible regulation mechanism. To test this idea, we carried out molecular dynamics (MD) simulations using OPLS-aa full-atom force-field [64] and the TIP4P explicit solvent model [63]. After 10 ns equilibration at 300 K, started from the 1.0 Å resolution crystal structure including calcium ions (pdb code 1EXR) [138], the required ensemble was extracted from a subsequent 78 ns simulation run. Additionally, the structure of the calcium-free calmodulin was modelled by equilibrating the same crystal structure in absence of Ca for 80 ns. The good agreement of CaM structures with the calcium-free NMR ensemble (pdb code 1CFC) [139] corroborates that the relaxation was complete. A subsequent 88 ns simulation served to generate an equilibrium ensemble. From both ensembles, entropies were calculated according to the method described above.

FCA was conducted on the first 2500 modes of both systems (Fig. 4.2, indicated as grey-shaded area), whereas the remaining modes were considered sufficiently close to harmonic to allow a simple QH-estimate of their entropic contributions. For the calcium-bound state, sufficiently small subspaces $d_s < 40$ were obtained, and density estimation

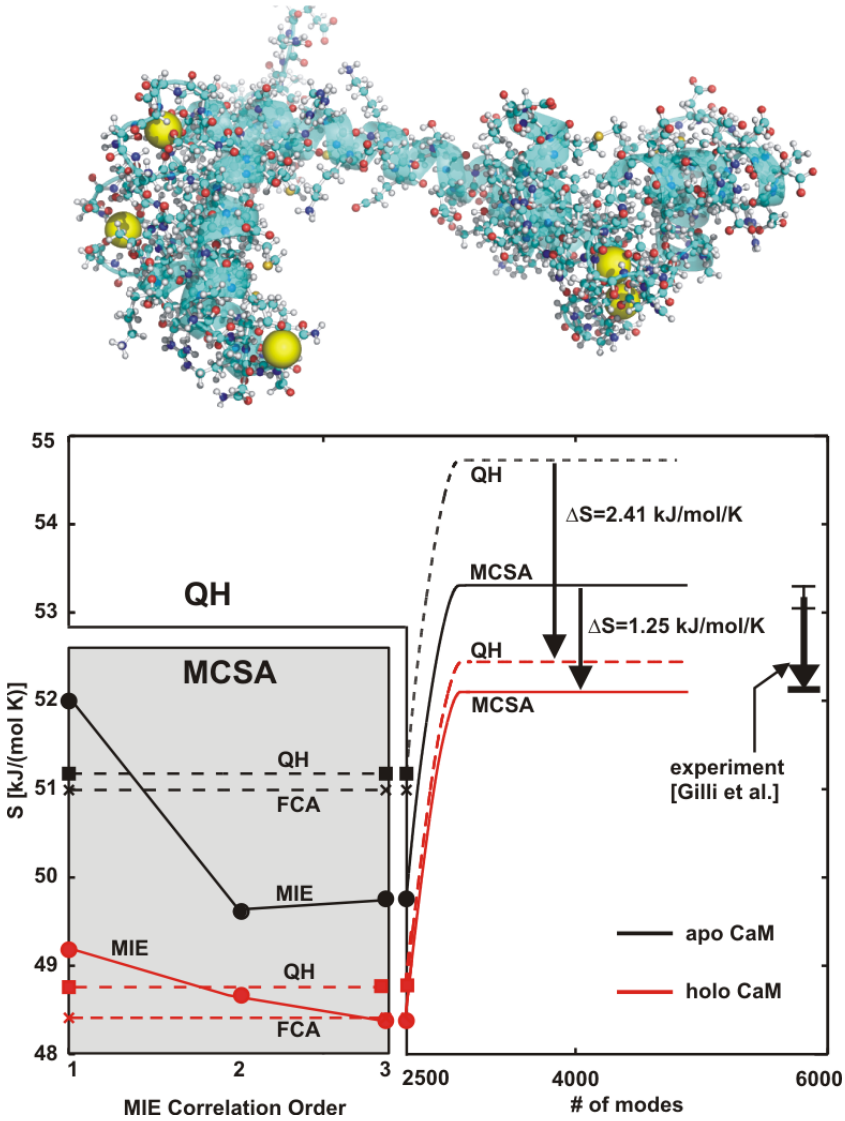

Figure 4.2: Minimally coupled subspace approach (MCSA) entropies for holo (red) and apo (black) calmodulin (CaM) and comparison with the respective QH estimates. The MCSA scheme was applied to the first 2500 modes (grey area). FCA: Full correlation analysis was carried out for the first 2500 modes, correlated modes were clustered as described in the text, and non-parametric density estimates were obtained. MIE: Additionally, MIEs (circles) were applied to this subset for each oversized cluster. QH: For comparison, the QH estimate is also given. White area: For the remaining high-frequency modes, the QH estimates were used and added to the MCSA estimates of the first 2500 modes. QH: QH-estimate, FCA: sum of density estimates of each minimally coupled subspace.

was carried out on each of them, thus lowering the total entropy by $0.4\,\mathrm{kJ\,(mol\ K)^{-1}}$ compared to the QH estimate. If, nevertheless, MIE ($h_s = 12$) is applied, a similar behaviour as for the smaller test systems discussed above is seen. In particular, also for CaM the fact that the 3rd order MIE recovers the FCA value confirms that truncation after the 3rd term provides accurate entropies also for larger systems. For the calcium-free state, in contrast, one oversized cluster of 96 modes remains; accordingly, and as expected, MIE improved the result further by $1.2\,\mathrm{kJ\,(mol\ K)^{-1}}$, which is $1.5\,\mathrm{kJ\,(mol\ K)^{-1}}$ lower than the QH approximation.

Both QH and MCSA consistently predicted a marked entropy drop upon calcium binding. However, MCSA yields a markedly smaller entropy drop of $1.25\,\mathrm{kJ\,(mol\ K)^{-1}}$ compared to the QH value of $2.41\,\mathrm{kJ\,(mol\ K)^{-1}}$, thus underscoring that inclusion of both non-linear and higher-order correlations are crucial.

To compare to experiment, we consider calorimetric experiments that report [140] an entropic gain of $387 \pm 30\,\mathrm{J\,(mol\ K)^{-1}}$ upon binding of the four calcium ions. This value includes calcium desolvation and binding contributions of $404 \pm 2\,\mathrm{J\,(mol\ K)^{-1}}$ leaving a nearly zero contribution of $-17 \pm 38\,\mathrm{J\,(mol\ K)^{-1}}$ due to protein structure and dynamics. To compare this value with our calculations, the molecular surface change of $\Delta\mathrm{MS} = -9.85\,\mathrm{nm^2}$ needs to be considered, which implies an entropy change [141, 142] of roughly [143] $\gamma = 96.3\,\mathrm{J\,(mol\ K)^{-1}nm^{-2}}$, i.e. $\Delta S_{\mathrm{solv}} = 950 \pm 240\,\mathrm{J\,(mol\ K)^{-1}}$, where the large error accounts for the large surface fluctuations observed in the simulations. From our MCSA result, $-280 \pm 240\,\mathrm{J\,(mol\,K)^{-1}} \approx 0$ is obtained, i.e. a nearly complete entropy cancellation within the error, in agreement with the experiment. This result suggests that calcium binding causes a substantial but free-energy-neutral transfer of configurational entropy into solvent entropy, thereby activating CaM for substrate binding. Quite a different picture would be provided by the QH approximation, which yields an entropy change of $-1460 \pm 240\,\mathrm{J\,(mol\,K)^{-1}}$, thereby clearly overestimating the effect of entropy reduction.

From a different perspective, correcting the calorimentric values by the solvent contributions, an experimental range of $-970 \pm 280\,\mathrm{J\,(mol\,K)^{-1}}$ is obtained (Fig. 4.2 right), which favourably compares with the MCSA result of $-1250\,\mathrm{J\,(mol\,K)^{-1}}$ but not with the QH value of$-2410\,\mathrm{J\,(mol\,K)^{-1}}$. A similar factor two overestimation of the QH approximation is seen for insulin between different protonation states [144].

Conclusion

We conclude that inclusion of non-linear and higher-order correlations as well as of anharmonicities is indispensable for computation of macromolecular configurational entropies. The method developed here circumvents the curse of dimensionality and accounts for these effects even for the huge configurational space spanned by proteins. For calmodulin, our method proved sufficiently accurate to correctly describe the fine-tuned entropy balance that governs its physiological activation. While the MIE expansion up to third order proved accurate in the studied systems, convergence issues are discussed in more detail in Chapter 6. FCA dominates the computational costs, which become demanding for over 2500 modes. Finally, our approach can serve to decompose the configurational entropy into individual contributions, e.g. from side chains or the backbone.

Chapter 5

5 Adaptive kernels for non-parametric estimation of configurational entropies of macromolecules

For accurate entropy calculations of biological macromolecules, it turns out that the hierarchical approach introduced in the last chapter requires accurate density estimates for up to 50 dimensions. However, current density estimators have been shown to yield accurate results for only up to 10 internal degrees of freedom, whereas failure was reported for a molecular system with 23 internal degrees of freedom[131]. In this chapter, a density estimator is therefore developed, which can be used in the hierachical approach.

To derive the density estimator, note that the intricate mix of stiff and soft degrees of freedom [145] typically found in macromolecules such as proteins yields a configurational density that is *threaded*, i.e., the density is extended in spatial directions that correspond to "soft" degrees of freedom, whereas it is confined to thin regions of space in the "stiff" directions. Accordingly, the established methods which average over isotropic regions in space will considerably blur this density thread locally perpendicular to the thread. It will be shown here that non-parametric entropy estimates are considerably improved by using *elliptic* kernels which are locally adapted to the threads in the configurational density. Because, still, the neccessary local averaging will give rise to remaining blurring, the latter will be corrected for by an empirical correction term parameterized by the average volume of the used kernel. Finally, generalizing the quasi-harmonic Schlitter formula [57], the quantum mechanical nature of the stiffest degrees of freedom will be

accounted for. To this aim, a size limit to each principal axis of the elliptic averaging kernel is employed here to ensure that each direction in space receives at least a blurring or uncertainty that prevents the occurrence of negative entropy contributions (see theory section).

5.1 Theory

5.1.1 Thermodynamic entropy

We first sketch the conceptual framework to clarify notation. The molecular dynamics of an isolated macromolecule with N atoms is described by the Hamiltonian

$$H(\mathbf{x},\mathbf{p}) = \frac{1}{2}\sum_{i=1}^{3N}\mathbf{p}_i^2 + V(\mathbf{x}),$$

where $\mathbf{x}$ and $\mathbf{p}$, are the mass-weighted $3N$-dimensional position and momentum vectors, respectively. Their cartesian components x_i and p_i are related to the non-mass weighted components $\tilde{x}_i$ and $\tilde{p}_i$ by $x_i = m_i^{1/2}\tilde{x}_i$ and $p_i = m_i^{-1/2}\tilde{p}_i$, respectively, where m_i denotes the mass of the respective atom. The entropy of the system is given by

$$S = \frac{\langle H\rangle}{T} + k_\mathrm{B}\ln Z,$$

where the angular brackets $\langle . \rangle$ denote an ensemble average, T the temperature, and Z is the classical partition function

$$Z = \frac{1}{h^{3N}}\int e^{-\beta H} d\mathbf{p}d\mathbf{x} \quad ,$$

with h being Planck's constant, and $\beta = 1/k_\mathrm{B}T$.

For conservative systems, Z separates into two dimensionless factors, $Z = Z_p Z_x$, with

$$\begin{aligned} Z_p &= \frac{1}{\kappa^{3N}}\left(\frac{2\pi}{\beta}\right)^{3N/2} \quad \text{and} \\ Z_x &= \frac{1}{\ell^{3N}}\int e^{-\beta V(\mathbf{x})} d\mathbf{x} \quad . \end{aligned}$$

Here, we have defined a convenient characteristic length $\ell = 1\,\mathrm{nm\,u}^{1/2}$ and, similarly,

$\kappa = h/\ell$. Accordingly, the entropy of the system falls into a kinetic and a configurational part,

$$S = S_p + S_c \quad ,$$

with

$$\begin{aligned} S_p &= 3Nk_\mathrm{B}\left(1 + \ln 2\pi/\kappa^2\beta\right)/2 \quad \text{and} \\ S_c &= \frac{\langle V(\mathbf{x})\rangle}{T} + k_\mathrm{B} \ln Z_x \, . \end{aligned} \tag{5.1}$$

Using the configurational density $\rho\left(\mathbf{x}\right) = \exp\left(-\beta V\left(\mathbf{x}\right)\right)/Z_x$ we express the configurational entropy as

$$S_c = -\frac{k_\mathrm{B}}{\ell^{3N}} \int \rho\left(\mathbf{x}\right) \ln \rho\left(\mathbf{x}\right) d\mathbf{x}, \tag{5.2}$$

which closely resembles Shannon's information entropy.

5.1.2 Quasi-harmonic approximation

To estimate S_c from a given (finite) ensemble of structures $\{\mathbf{x}\}$ obtained, e.g., from a molecular dynamics or Monte-Carlo simulation, the quasi-harmonic approximation is commonly used, where

$$\rho(\mathbf{x}) \approx \tilde{\rho}\left(\mathbf{x}\right) = \exp\left[-\frac{1}{2}(\mathbf{x} - \langle\mathbf{x}\rangle)^{\mathbf{T}} \mathbf{C}^{-1} \left(\mathbf{x} - \langle\mathbf{x}\rangle\right)\right] \quad ,$$

and $\mathbf{C}$ denotes the covariance matrix

$$\mathbf{C} = \left\langle \left(\mathbf{x} - \langle\mathbf{x}\rangle\right)\left(\mathbf{x} - \langle\mathbf{x}\rangle\right)^{\mathrm{T}} \right\rangle .$$

In this approximation the configurational density factorises,

$$\rho_\mathrm{QH}\left(\mathbf{y}\right) = \prod_{i=1}^{3N} \rho_i\left(y_i\right), \tag{5.3}$$

with marginal densities $\rho_i\left(y_i\right) \propto \exp\left(-\beta y_i^2/2\lambda_i\right)$, where $\mathbf{y} = \mathbf{T}\left(\mathbf{x} - \langle\mathbf{x}\rangle\right)$ are the principal coordinates derived from diagonalisation of $\mathbf{C}$, i.e. $\mathbf{T}^T\mathbf{C}\mathbf{T} = \mathrm{diag}\left(\lambda_1, \ldots, \lambda_{3N}\right)$. Using

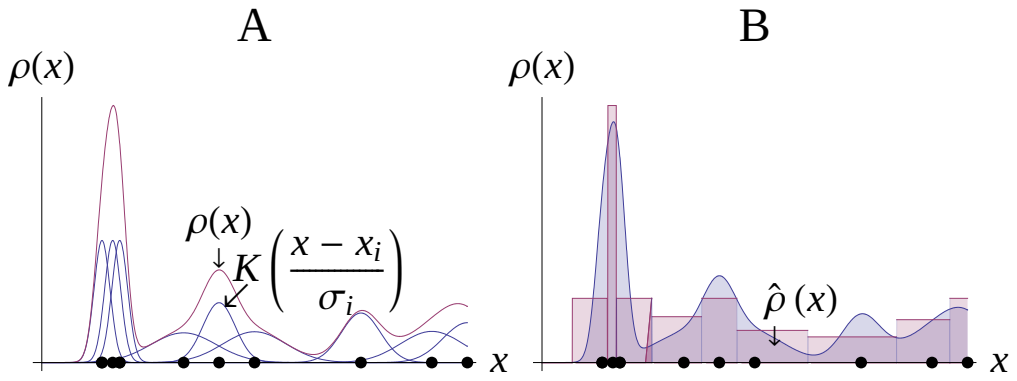

Figure 5.1: Principle of kernel density estimation with Gaussian kernels. Other kernels were also tested. A, the configurational space density $\rho(\mathbf{x})$ vs the local kernel approximation at sample points $\mathbf{x}_i$, indicated as black dots. B, the configurational space density is approximated as sum, $\hat{\rho}(\mathbf{x})$, of locally constant densities of Voronoi cells, $\hat{\rho}(\mathbf{x}_i)$, around $\mathbf{x}_i$.

the orthonormality of $\mathbf{T}$, and combining Eqs. 5.1 and 5.2, the quasi-harmonic entropy estimate,

$$S_{\mathrm{QH}} = \frac{1}{2} k_{\mathrm{B}} \sum_{i=1}^{3N} \ln \left[\frac{e^2 \lambda_i}{\beta \hbar^2} \right], \tag{5.4}$$

is obtained.

We note that this estimate holds only for the molecular frame, with rigid body motions properly removed. Eq. 5.4 has been elegantly generalised to account for the quantum mechanical nature of the vibration of stiff degrees of freedom [57],

$$S_{\mathrm{qm,harm}} = \frac{1}{2} k_{\mathrm{B}} \sum_{i=1}^{3N} \ln \left(1 + \frac{e^2}{\beta \hbar^2} \lambda_i \right). \tag{5.5}$$

This approximation, which will be used below, in particular avoids divergencies for $\lambda_i \to 0$, which is an artifact of the classical treatment, Eq. 5.4.

5.1.3 Locally adapted non-parametric entropy estimation

We will use this framework to develop a locally adapted non-parametric density estimation based on the k-nearest neighbour density estimate of n sample points $\{\mathbf{x}\}$ which we will apply to $d \leq 3N$ dimensions,

$$\hat{S}_k = \frac{k_\mathrm{B}}{n} \sum_{i=1}^{n} \ln \frac{n\pi^{d/2} r_{i,k}^d}{k\Gamma\left(\frac{1}{2}d+1\right)\ell^d} \quad . \tag{5.6}$$

Here, $r_{i,k}$ is the Euklidean distance between the sample point $\mathbf{x}_i$ and its k-nearest neigbour, and Γ denotes the Gamma function.

The k-NN estimator rests on the straightforward assumption of the local density at sample point $\mathbf{x}_i$

$$\rho_k(x_i) \approx \frac{k\ell^d}{nV_d\left(r_i(k)\right)}\,, \tag{5.7}$$

where $V_d\left(r_{i,k}\right)$ denotes the volume of the d-dimensional sphere with radius $r_i(k)$ which is chosen such that the d-dimensional sphere centered at $\mathbf{x}_i$ contains k sample points.

5.1.3.1 Soft degrees of freedom

A generalisation of Eq. 5.6 to arbitrary kernel functions K for given k and for each point $\mathbf{x}_i$ is obtained by requiring $\sigma_{i,k}$ to be chosen such that

$$k = \sum_{j=1}^{n} K\left(\frac{\mathbf{x}_i - \mathbf{x}_j}{\sigma_{i,k}}\right). \tag{5.8}$$

This yields a potentially improved local (smoothed) density estimate at $\mathbf{x}_i$,

$$\rho(\mathbf{x}_i) \approx \rho_k\left(\mathbf{x}_i\right) = \frac{k\ell^d}{n\sigma_{i,k}^d Z_d}, \tag{5.9}$$

where $Z_d = \ell^{-d}\int K\left(\mathbf{x}-\mathbf{x}_i\right)d\mathbf{x}$. The probability density distribution $\rho(\mathbf{x})$is then approximated as a sum of these local densities (Fig. 5.1 B),

$$\rho\left(\mathbf{x}\right) \approx \hat{\rho}\left(\mathbf{x}\right) := \sum_{i=1}^{n} \theta_i\left(\mathbf{x}-\mathbf{x}_i\right)\rho_k\left(\mathbf{x}_i\right),$$

where the Voronoi function θ_i is unity for all $\mathbf{x}$ that are closer to $\mathbf{x}_i$ than to any other $\mathbf{x}_j$, $i \neq j$, and zero otherwise. This approximation dissects the entropy integral, Eq. 5.2,

$$\begin{aligned} S_c &\approx -k_\mathrm{B} \int \hat{\rho}\left(\mathbf{x}\right) \ln \hat{\rho}\left(\mathbf{x}\right) \, \mathrm{d}\mathbf{x} \\ &\approx -k_\mathrm{B} \sum_{i=1}^{n} \ln \hat{\rho}\left(\mathbf{x}_i\right) \, , \end{aligned}$$

$$\int \rho\left(\mathbf{x}\right) \log \rho\left(\mathbf{x}\right) \mathrm{d}\mathbf{x} = \rho_k\left(\mathbf{x}\right) \log \rho_k\left(\mathbf{x}\right) / \rho_k\left(\mathbf{x}\right) / n$$

where the volume of the Voronoi cells was assumed to be given by the inverse local density.

Accordingly, the entropy contribution from the d considered degrees of freedom is estimated by

$$\hat{S}_{\text{g-NN}} = \frac{k_\mathrm{B}}{n} \sum_{i=1}^{n} \log \frac{n Z_d \sigma_{i,k}^d}{k \ell^d} + \frac{d}{3N} S_p, \tag{5.10}$$

which generalises Eq. 5.6 and, thus, may be denoted in the following as g-NN. The k-NN formula is recovered by using the rectangular and isotropic kernel function

$$K_{k\text{-NN}}\left(\frac{\mathbf{x}_i - \mathbf{x}_j}{\sigma_{i,k}}\right) = \begin{cases} 1 & \|\mathbf{x}_i - \mathbf{x}_j\| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

and setting $\sigma_{i,k} = r_{i,k}$, since for this kernel $Z_d = V_d(1)$.

The Gaussian function framework underlying the quasi-harmonic approach sketched above suggests to use instead a Gaussian kernel [79]

$$K\left(x\right) = \exp\left(-\left\|\mathbf{x}^2\right\|\right). \tag{5.11}$$

However, as illustrated in Fig. 5.2, for such isotropic kernels, the approximation Eq. 5.10 tends to fail for the thin "threads" typically encountered for the configurational density of biological macromolecules, i.e. where the width of the "thread" is smaller than the average distance between adjacent sample points (A). To address this issue, we propose to locally adapt the kernel function to the stiff degrees of freedom using for each $\mathbf{x}_i$

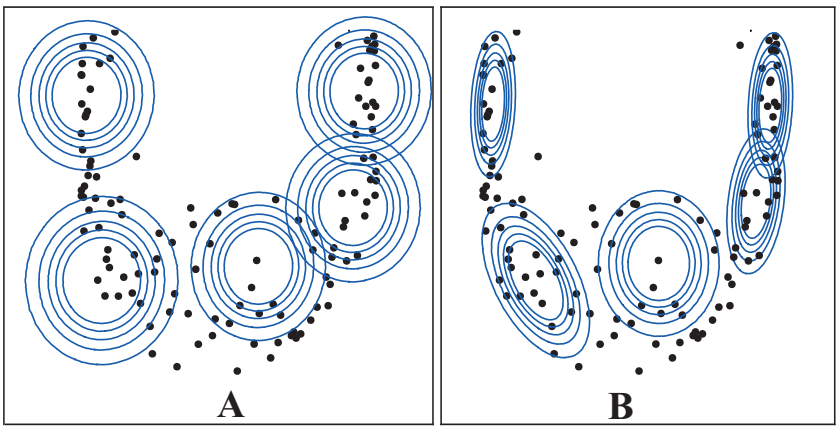

Figure 5.2: Isotropic vs. locally adapted anisotropic kernels. For threaded configurational space densities (sampled by the black dots), isotropic kernels (A) fail to provide accurate density approximations. In contrast, locally adapted anisotropic kernels (B) improve the approximation particularly perpendicular to the threads.

anisotropic Gaussian kernels (B),

$$K_{\text{loc,gauss}}(\mathbf{x}) = \exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{x}_i)^T \mathbf{A}(\mathbf{x}-\mathbf{x}_i)\right], \tag{5.12}$$

with

$$\mathbf{A}^{-1} = \mathbf{C} = \frac{1}{k}\sum_{i=1}^{k}(\mathbf{x}-\mathbf{x}_i)(\mathbf{x}-\mathbf{x}_i)^T,$$

given by the covariance matrix $\mathbf{C}$ and the sum runs over the k nearest neighbours of $\mathbf{x}_i$. Elliptic hard kernels are obtained analogously

$$K_{\text{loc,ell}}(\mathbf{x}) = \begin{cases} 1 & \left\|(\mathbf{x}-\mathbf{x}_i)^T \mathbf{A}(\mathbf{x}-\mathbf{x}_i)\right\| \leq 1 \\ 0 & \text{otherwise} \end{cases}. \tag{5.13}$$

5.1.3.2 Stiff degrees of freedom — Quantum correction

The locally adapted anisotropic Gaussian kernels naturally define stiff degrees of freedom in a canonical way. For those, it is possible to obtain physically appropriate quantum corrections. The classical treatment of the stiff degrees of freedom yields unphysical entropies by allowing unlimited sharpness of the “threads”. In contrast, the Schlitter

formula, Eq. 5.5, effectively requires stiff degrees of freedom to assume density distributions of a minimal width $\sigma_{\mathrm{qm}} = \beta\hbar^2/e^2$. Schlitter's treatment can be generalised to arbitrary densities by computing the entropy from

$$\rho_\sigma(\mathbf{x}) = \int \rho(\mathbf{x}) K(\mathbf{y} - \mathbf{x}/\sigma_{\mathrm{qm}}) dy, \tag{5.14}$$

rather than from the true density $\rho(\mathbf{x})$, where $K(\mathbf{y})$ denotes a d-variate smoothing kernel of unit width. Applying this convolution to the quasi-harmonic density Eq. 6.9 with a Gaussian smoothing kernel K (Eq. 5.11) of width σ_{qm} yields a smeared out multivariate Gaussian density whose widths are given by $\lambda_i + \sigma_{\mathrm{qm}}$. Accordingly, one retrieves Schlitter's formula, Eq. 5.5.

In contrast to Schlitter's treatment, our generalisation also holds, e.g. for a bimodal density of two non-overlapping Gaussian functions,

$$\rho(\mathbf{x}) = \frac{\ell}{2\varepsilon}\sqrt{\frac{1}{2\pi}}\left[\exp\left(-\frac{(\mathbf{x}-\mathbf{x}_0)^2}{2\varepsilon^2}\right) + \exp\left(-\frac{(\mathbf{x}+\mathbf{x}_0)^2}{2\varepsilon^2}\right)\right],$$

with $x_0 \gg \varepsilon$. Convolution according to Eq. 5.14 yields the entropy $S(\varepsilon) = k_{\mathrm{B}} \ln 2$ for $\varepsilon \to 0$, demonstrating that this generalisation yields meaningful results beyond the quasi-harmonic approximation.

The smoothing inherent to the locally adapted kernel density estimation simultaneously provides an approximate and canonical treatment of the stiff degrees of freedom. Accordingly, we restrict the width $\sigma_{j,i}$ of the g-NN smoothing kernel i in the local direction j to σ_{qm} by $\sigma_{j,i,\mathrm{qm}} = \max(\sigma_{i,j}, \sigma_{\mathrm{qm}})$, where the minimal width is set to

$$\sigma_{\mathrm{qm}} = \frac{\pi\hbar^2\beta}{eZ_d^{2/d}},$$

such that if $\sigma_{i,k} = \sigma_{\mathrm{qm}}$,

$$\hat{S}_{\text{g-NN}} = \frac{k_{\mathrm{B}}}{n}\sum_{i=1}^{n} \ln\frac{n}{k} \quad .$$

Thus, $\hat{S}_{\text{g-NN}} \to 0$ for $k \to n$, in the special case that the density is concentrated within a single region below the quantum resolution limit and $\hat{S}_{\text{g-NN}} = k_{\mathrm{B}} \log M$ if the density is distributed equally between M non-overlapping areas in phase space that are all narrower than the quantum resolution limit.

5.1.3.3 Empirical Smoothing Correction

A correction for the smoothing of the configurational density implied by the convolution with the g-NN Gaussian kernel functions within our Gaussian nearest neighbour approximation was derived from Eq. 5.4. Inverting Eq. 5.4 yields an effective width λ_{eff} of a hypothetical isotropic d-dimensional Gaussian density function that would correspond to a given entropy S,

$$\lambda_{\text{eff,qh}} = \frac{\hbar^2\beta}{e^2} \exp\left(\frac{2}{dk_{\text{B}}}S\right),$$

A deviation of the entropy estimated by the g-NN method S_{gNN} from the true entropy S can thus be expressed as $\Delta\lambda = \lambda_{\text{eff,qh}} - \lambda_{\text{eff,gNN}}$. Convolution of a Gaussian density with a kernel function of width σ results in a density of width $\lambda_{\text{eff,qh}} + \sigma$. We therefore assume a functional relationship of $\Delta\lambda$ with the average kernel width $\bar{\sigma}$. To test this hypothesis in high-dimensional space, we generated multi-variate Gaussian densities with varying widths. The corresponding entropy was computed analytically and compared to S_{gNN}. We found for the used set of test-functions a good approximation to $\Delta\lambda$ by the linear relationship $\Delta\lambda = m \max(0, \bar{\sigma} - \sigma_{\text{qm}}) + b$, where

$$\begin{aligned} m &= -1.015 \times 10^{-3} d + 0.079 \\ b &= -5.4 \times 10^{-8} d + 8.5 \times 10^{-7}. \end{aligned}$$

Thus, a corrected entropy is given by

$$S_{\text{corr}} = \frac{dk_{\text{B}}}{2} \ln\left[\frac{e^2}{\hbar^2\beta}\left(\lambda_{\text{eff,gNN}} + \Delta\lambda\right)\right].$$

Because of all functions with given variance, the Gauss function has the largest entropy, S_{corr} is guaranteed to provide an upper bound for the true entropy.

5.2 Methods

5.2.1 Simulation setup

The test systems butane to decane and dialanine, which were compared with a thermodynamic integration reference (see below) were set up as follows. Force-field parameterizations were obtained from the Dundee Prodrug server [146] based on the GROMOS

united-atom force field [65]. Stochastic Dynamics simulations were performed using the molecular simulations package GROMACS [113] in vacuo with friction constant γ set to 10, dielectric constant $\epsilon = 1$, integration step size of 0.0005 ps and no bond constraints. Positional restraints were applied to three adjacent terminal heavy atoms.

The coldshock protein (protein database entry 1CSP) was simulated using the OPLS all atom force field [64] and periodic boundary conditions. NpT ensembles were simulated, with the protein and solvent coupled separately to a 300-K heat bath ($\tau = 0.1$ ps) [69]. The systems were isotropically coupled to a pressure bath at 1 bar ($\tau = 1.0$ ps) [69]. Application of the Lincs [31] and Settle[30] algorithms allowed for an integration time step of 2 fs. Short-range electrostatics and Lennard–Jones interactions were calculated within a cut-off of 1.0 nm, and the neighbour list was updated every 10 steps. The particle mesh Ewald (PME) method was used for the long-range electrostatic interactions [72], with a grid spacing of 0.12 nm.

5.2.2 Reference entropies by Thermodynamic Integration

Absolute free energies for the test systems butane to decane and dialanin were calculated by thermodynamic integration (TI). No reference values were obtained for the cold shock protein (CSP), because the system is too large for the TI to converge. Simulation parameters were as discussed above. The TI scheme we have chosen to obtain the Helmholtz free energy A of the fully interacting particles consists of two phases (also shown in Fig. 5.3). Harmonic position restraints with a force constant $k = 25000\,\mathrm{kJ/mol/nm^2}$ were slowly switched on for each atom in the first phase, and in the second phase all force-field components were gradually switched off. The system then consisted of non-interacting dummy particles with mass m oscillating in their respective harmonic position restraint potentials, i.e.,

$$V = \frac{1}{2}k\sum_{j=1}^{N}(\mathbf{x}-\mathbf{x}_j)^2\,.$$

The free energy of this harmonic system can be obtained analytically,

$$A_0 \;=\; -\beta^{-1}\frac{3}{2}\sum_{j=1}^{N}\left[\log\left(\frac{1}{\hbar^2\beta^2\tilde{k}_j}\right)\right]$$

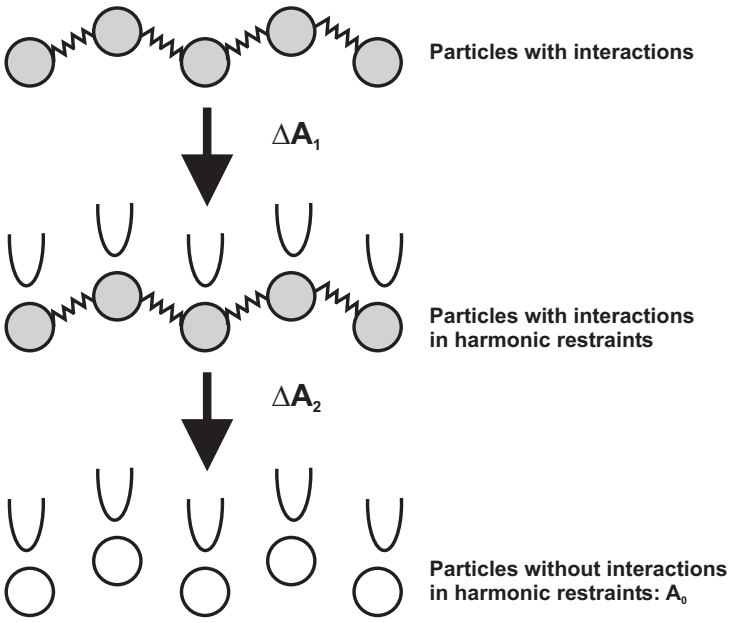

Figure 5.3: Thermodynamic integration scheme used for obtaining reference entropies. Grey circles depict interacting particles, white circles non-interacting particles, zig-zags represent chemical bonds, and local harmonic potentials are sketched as parabolas. Partial charges were removed in a separate step not illustrated here. A similar scheme has been used by Tyka et al [130].

where $k = \tilde{k}/m_j$ denotes the mass-weighted force constant. Hence, the thermodynamic integration yields the absolute free energy

$$A \quad = A_0 \quad -\Delta A_2 - \Delta A_1$$

and the entropy by $S = (A - \langle V \rangle)/T$, where $\langle V \rangle$ denotes the ensemble average of the potential energy.

For the TI between the systems given by V_s (start) and V_f (end), 18 intermediate steps $V_i(\lambda) = \lambda V_s + (1 - \lambda) V_f$, $i = 1, \ldots, 18$ were used, and the intermediate values of $\lambda_i = 0$, 1e-6, 5e-6, 1e-5, 5e-4, 1e-4, 1e-3, 1e-2, 2e-2, 3e-2, 5e-2, 7e-2, 9e-2, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1 were distributed unevenly to obtain approximately balanced ΔA_i values. For each value of λ a trajectory of 125 ns was generated.

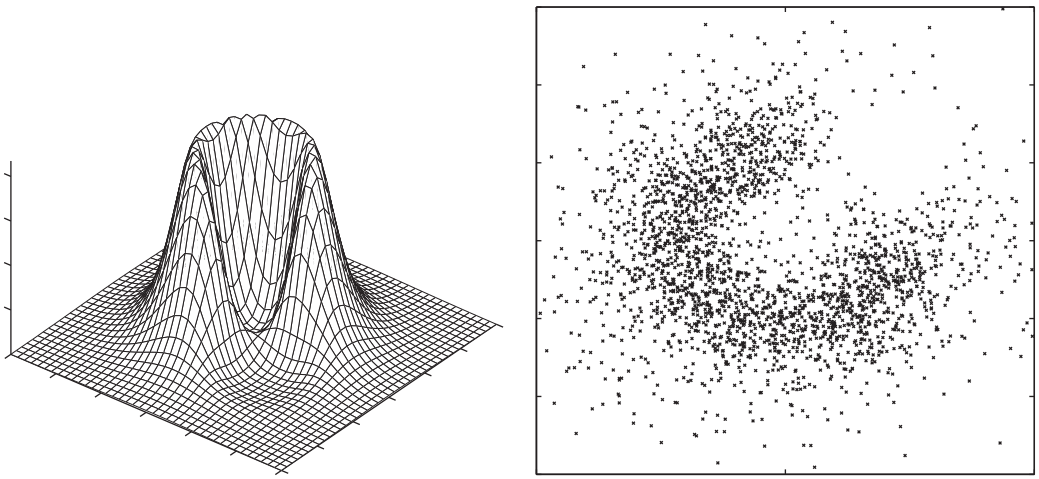

Figure 5.4: Left, synthetic test 2D density whose entropy, $97.8\,\mathrm{J\,K^{-1}\,mol^{-1}}$, was computed numerically by integrating on a grid. Right, 5000 points drawn according to this density function were used to estimate an entropy of $97.2\,\mathrm{J\,K^{-1}\,mol^{-1}}$.

5.2.3 Efficient implementation

To compute the entropy via g-NN (Eq. 5.10) as described is computationally more expensive than for k-NN (Eq. 5.6), because to solve Eq. (5.8) until convergence requires a large number of iterations. Note, however, that the density estimate Eq. 5.9 is for all practical purposes invariant for small changes of k, because both k and $\sigma_{i,k}$ appear in Eq. 5.9 and scale similarly. Accordingly, convergence of k to 10% was considered sufficient to achieve accurate results, thus drastically reducing computational cost to a level similar to k-NN.

The kd-tree implementation of the nearest neighbour library ANN[147] has been used for fast look-up of the neighbouring sample points $\mathbf{x_j}$ with $r_{ij} < r_c$.

5.3 Results and Discussion

5.3.1 Example: Simple density distributions

In a first step, our non-parametric entropy estimation was tested with synthetic probability density distributions whose entropy is accessible either analytically (e.g., Gaussians) or through grid-based numerical approximation (e.g., densities in 2- and 3- dimensional space). Figure 5.4 shows one of several test densities which were designed to exhibit

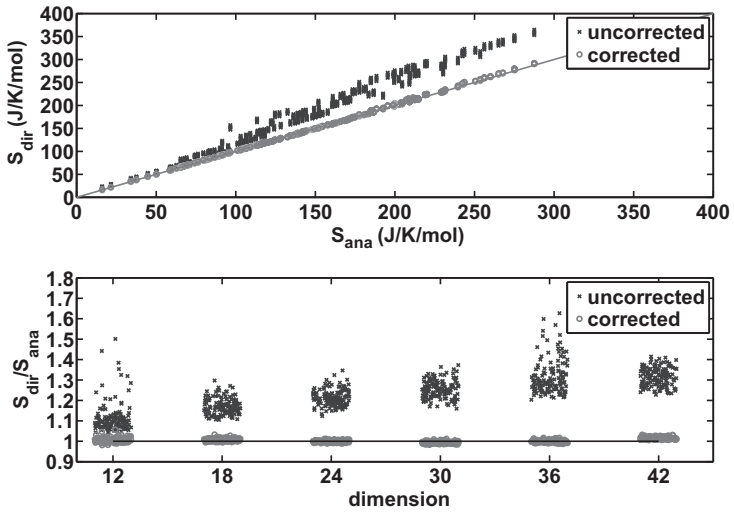

Figure 5.5: Effect of the empirical smoothing correction described in the text. Top: Systematic overestimation seen as deviations from the diagonal. Black crosses: uncorrected values; grey circles: corrected values. Bottom: The same data as a function of dimensionality.

typical features, e.g. a curved "thread", also seen in the configurational densities of macromolecules. All reference entropies were reproduced by the estimator to within $1\,\mathrm{J\,K^{-1}\,mol^{-1}}$ or better (results not shown). A similar test density was studied in three dimensions, as well as 42-dimensional checkerboard-like densities.

As a more challenging, high-dimensional test, and aiming to assess the empirical smoothing correction derived above, we considered Gaussian densities ranging from 12 to 42 dimensions. Fig. 5.5a shows reference entropies S_{ref} versus values obtained from our density estimator, both with (black) and without (grey) empirical smoothing correction. Fig. 5.5b shows the same data as a function of dimensionality. As can be seen, the correction improves the obtained entropy values considerably and provides accurate values independent of dimensionality. In contrast, the uncorrected values systematically overestimate the entropy with increasing dimension. This overestimation is due to the smoothing effect described in Section 5.1.3.3, which, due to increased surface effects aggravates with increasing dimensionality.

System	N	A_{TI}	H	S_{TI}	S_{dir}	S_{nc}	S_{iso}	S_{FCA}	*clust*	S_{QH}
Butane	4	-34	39.8	184	187	191	201	194	5	217
Pentane	5	-48.4	51.3	249	250	256	281	255	8	307
Hexane	6	-63.6	60.7	311	317	327	373	321	11	404
Heptane	7	-77.1	72.5	374	387	403	476	393	13	499
Octane	8	-95.4	81.3	441	448	469	575	455	15	617
Nonane	9	-111.1	89.5	501	525	553	687	529	19	711
Decane	10	-128.6	101	540	592	598	827	624	21	808
Dialanine	15	-66.4	144.4	527	550	610	—	549	32	893
CSPbb	201	—	—	—	3081			2717	43	3081

Table 5.1: Detailed test results for all alkanes ethane – decane, dialanine, 12-residue beta turn and the backbone of a coldshock protein (pdb code 1CSP). Herein: A_{TI}: absolute free energy from *TI* (in kJ/mol); H internal energy (in kJ/mol); S_{TI}: absolute configurational entropy obtained from *TI* (in J/mol/K); S_{dir}: direct adaptive kernel density enstimation without any subspace clustering; S_{nc}: like S_{dir} but without smoothing correction (cf. methods); S_{iso}: direct isotropic kernel density estimation without any subspace clustering; S_{FCA}: sum of density esimates after subspace clustering; *clust*: size of largest cluster; S_{QH}: entropy estimate according to quasi-harmonic approximation.

5.3.2 Example: Alkanes

We next applied our estimation method to a number of more realistic molecular test systems. Here, the accuracy of the estimate may be affected by two sampling effects; first, insufficient simulation sampling due to unvisited configurational space regions, second, locally too sparse sampling, which affects the accuracy of the NN density estimates. These two sampling effects are largely independent; whereas the sparse sampling problem, also called the "curse of dimensionality" [132], aggravates with increasing dimensionality and is largely independent of the particular system studied, the simulation sampling problem will depend on the size of the accessible configurational space as well as the slowest relaxation times of that system. Consider, for example, a single harmonic well. Whereas there is no simulation sampling problem for high-dimensional wells, NN-density estimations will inevitably suffer from sparse sampling for high-dimensional wells — a problem which, due to the regularisation assumptions, is much less pronounced for the quasi-harmonic approximation.

As test systems we chose (i) n-alkanes ranging from butane to decane, as a model of

protein sidechain behavior, and (ii) dialanine, as a minimal model for a protein backbone. The flexible worm-like shape of alkanes is likely to aggravate the two sampling problems discussed above sufficiently to explore the limits of our density estimator.

To obtain reference entropy values for these systems, a thermodynamic integration scheme was used which gradually perturbed the systems towards an analytically tractable state consisting of non-interacting particles in harmonic wells, as described in methods, Section 5.2.2. In each case, convergence was reached (data not shown).

The seven n-alkanes ranging from butane to decane were described by a united atom force-field, such that the systems comprised four to ten atoms, respectively. Fig. 5.6 shows the entropies obtained with density estimation for these systems (cf. also Table 5.1) as well as the reference obtained by thermodynamic integration and the QH entropy for comparison. Each of the listed density estimates has been obtained by averaging five independent trajectories. As can be seen, the QH entropy (dashed line) overestimates the configurational entropy by a wide margin for all systems ($15 - 50\%$). The adaptive kernel density estimation, in contrast, yielded considerably improved results. For butane to octane, a deviation below 3 % from reference was achieved. For the largest alkanes considered, nonane and decane, a slightly reduced accuracy of 5% and 14%, respectively, was obtained.

It is instructive to consider the density estimation results for the conventional isotropic kernels, S_{iso}, also listed in Table 5.1. Here, apparently due to the "curse of dimensionality" [132], drastic overestimates are seen for all alkanes. For butane (12 degrees of freedom) the entropy is overestimated by 9%, and for decane performs even worse than the QH approximation.

This result shows that it is the locally adapted anisotropic kernels (ellipsoids rather than spheres, whose principal axes are determined from the local density by a principal component analysis, c.f. Methods), which provide the observed accuracy. Anisotropic kernels according to Eq. 5.13 rather than according to Eq. 5.12 turned out to yield more accurate results (data for the latter not shown).

As can also be seen in the figure, the smoothing correction, which reverts the inevitable blurring during the averaging process of the density estimation (cf. Methods), markedly improved the density estimates.

The inset of Fig. 5.6 highlights the convergence behaviour of the locally adapted kernel estimates for alkanes with increasing simulation length. Clear convergence is seen for all alkanes up to octane, i.e. up to 24-dimensional densities. Whereas for these alkanes,

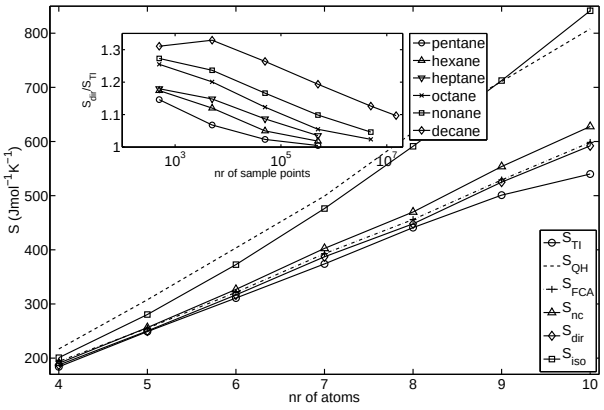

Figure 5.6: Entropies for the n-alcanes from butane (N=4) to decane (N=10) at 400 K. The entropies shown correspond to the columns of Table 5.1 and are defined in the respective caption. The errors of respectively $S_{\mathrm{dir}}, S_{\mathrm{nc}}, S_{\mathrm{iso}}$ and S_{FCA} were below 0.23 J mol^{-1}K^{-1} in all cases and not shown in the figure, since they are smaller than the symbols. The inset shows the same data as the relative derivation to the reference TI value, indicating convergence behaviour.

500.000 sample points sufficed for an accurate density estimate, 5.000.000 were required for nonane, and even more for decane, where sufficient sampling was not reached. In all cases, four sample points per picosecond simulation time were recorded. Although it is unclear at this point whether this lack of convergence is due to insufficient sampling of the simulation or the "curse of dimensionality" due to sparse sampling described above, the high flexibility of decane suggests the former. In the next section, we will address this question by considering a less flexible, but larger system.

5.3.3 Example: Di-alanine

Dialanine, as a minimal model of a protein backbone comprising 45 dimensions (15 united atoms), served as our largest test system. The molecule was simulated with implicit solvent at temperatures 300 K and 400 K using a united-atom force field. Fig. 5.7 shows entropies S_{dir} and S_{QH} obtained for simulation lengths ranging from 100 ps to 500 ns. The TI reference entropy is shown as a horizontal line. Similarly as seen above for the alkanes, the QH entropy S_{QH} overestimates the TI reference. Interestingly, and

in contrast to the other estimates, the QH-estimate increases with the length of the simulation. As a consequence, the longest trajectory (500 ns) yields the most inaccurate result with an error of 250%. We attribute this peculiar behaviour to a decreasing quality of the harmonic approximation with increasing complexity of the sampled configurational space. In contrast, the entropy obtained with density estimation S_{dir} converges towards the correct value with increasing simulation length. For the 500 ns simulation, the TI reference is reached to within 4.0% for 400 K and 18% for 300 K, respectively.

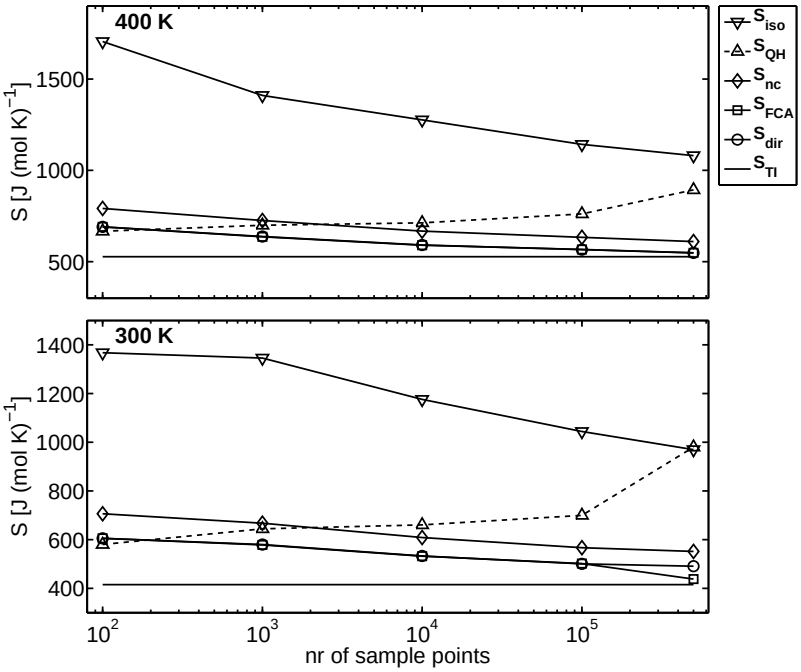

Figure 5.7: Calculated dialanine entropies vs. number of used sample points. The entropy labels shown are those defined in Table 5.1.

Again, density estimates from isotropic kernels provided even worse results than the QH estimate in all cases. Further, compared to the alkanes, the smoothing correction improved density estimates to an even larger extent; and the uncorrected estimates overestimated the reference by 16% for 400 K and 33% for 300 K, respectively.

This result corroborates the previous observation that the achieved accuracy was mainly due to the locally adapted anisotropic kernels. Moreover, even though the dimension

of the configurational space of dialanine is almost twice as large as that of nonane, the obtained accuracy is similar despite the fact that ten times less sample points were used for dialanine. Apparently, the curse of dimensionality is not yet severe here, thus underscoring the robustness of the locally adapted anisotropic kernels even for high dimensionality. A second conclusion is that, as expected, lack of simulation sampling limited the accuracy of the decane entropy estimates. This interpretation is further supported by the fact that more accurate estimates are obtained for the high temperature simulations, which are likely to provide enhanced simulation sampling, but sparser local sampling. In this sense, alkanes proved to be particular hard test systems.

5.3.4 Application: Coldshock protein

The previous sections established that locally adapted anisotropic kernel density estimation is very well feasible in configurational space of up to 45 dimensions. Although this is still far from the high dimensionality typically found for biological macromolecules, it holds the promise to provide the missing building block for a reliable application of the independent subspace framework (MCSA) derived in the previous chapter. Briefly, full correlation analysis [97] (FCA) is used to obtain nearly independent orthogonal configurational subspaces, which allow to approximately factorise the configurational space density. Subsequently, locally adapted anisotropic kernels are used to estimate the entropy contributions for each of these subspaces.

To test the feasibility and accuracy of this approach, we applied this framework to the coldshock protein, a 67 residue soluble protein (protein database entry 1CSP) which was simulated with the OPLS full atom force-field [64] in explicit solvent. Of particular interest was the question of how many (soft) degrees of freedom should be subjected to FCA and to the locally adapted anisotropic kernel density estimation, leaving the remaining (stiff) degrees of freedom to the Schlitter QH approximation. For simplicity and computational efficiency, only backbone contributions to the configurational entropy were considered. FCA was carried out on the first 100, 200, ..., 500 modes of the configurational ensemble, and nearly non-correlated subspaces were defined as decribed in the previous chapter. Our locally adapted anisotropic kernel estimate was subsequently applied to each of the resulting independent subspaces yielding an improved entropy estimate S_{FCA}. The remaining modes were considered via the Schlitter QH approximation, Eq. 5.5. The sum of these two yields the desired entropy estimate S_{tot}, also shown in Figure 5.8. As

expected, no TI entropy reference could be obtained due to lack of convergence; hence, we here resort to comparison with the QH estimate of the total configurational space density $S_{\text{tot, QH}}$.

Fig. 5.8 shows that the combination of FCA and locally adapted anisotropic kernels improves on the QH estimate in all cases. In particular, a marked improvement on the QH estimate is seen for the locally adapted anisotropic kernel estimate, which shows that the neglect of non-linear and higher-order correlations, and of anharmonicities by the established QH method implies an overestimate of the entropy by at least 14% (QH: $3081\,\text{JK}^{-1}\text{mol}^{-1}$ vs. S_{FCA}: $2710\,\text{J K}^{-1}\text{mol}^{-1}$). For the coldshock protein, further, 400 modes need to be subjected to he independent subspace approach for converged results, such that only 200 modes are properly described by the simple QH approximation.

This empirical distinction between stiff and soft modes is nicely reflected in the eigenvalue spectrum shown in the inset of Fig. 5.8. Whereas eigenvalues of modes 1 to 400 decrease smoothly, an abrupt drop is seen at mode 400. Similar observations for calmodulin (previous chapter) suggests that such abrupt drop can be used to identify the stiff modes that can be subjected to the QH approximation. Our results also suggest that, generally, a large fraction of degrees of freedom needs to be subjected to MCSA to obtain reliable estimates.

5.4 Conclusions

We have introduced a kernel based density estimation method and showed that it provides accurate results in even the high-dimensional and quite complex configurational space density generated by the dynamics of biological macromolecules. Whereas established k-nearest neighbour estimators have been reported to fail to converge for 23-dimensional configurational space with 15 million sample points, the locally adapted anisotropic kernels developed here provide robust results in up to 45 dimensions with only $500,000$ sample points. We attribute this improvement to the occurence of typically sparsely sampled configurational density "threads" ubiquitous in protein dynamics, to which, apparently, our locally multivariate approach provides an enhanced approximation.

Used within the framework of the minimally coupled subspace approach presented in the previous chapter, this density estimator serves to overcome the three limitations inherent in the conventional quasi-harmonic approximation, i.e. neglect of non-linear and higher order correlations, of anharmonicity, and of multi-modality.

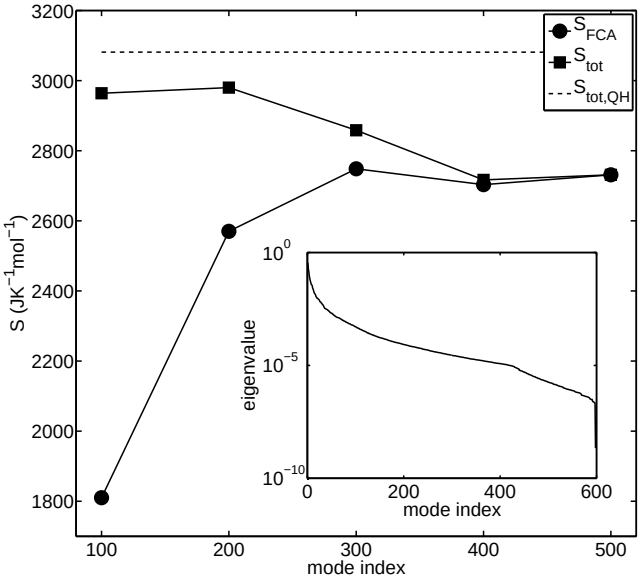

Figure 5.8: Estimated entropies for the coldshock protein (1CSP, backbone only). The total estimate, S_{tot}, as a function of MCSA subspace size ($100, \ldots, 500$ modes), was calculated as sum of S_{FCA}, the estimate obtained by application of locally adapted anisotropic kernels to independent subspaces obtained by FCA, and the QH estimate for the remaining modes, as described in the text; S_{PCA} : cumulative QH estimate; $S_{\mathrm{QH,tot}}$: global QH reference. For detailed values cf. Table 5.1. The inset shows the PCA eigenvalues as a function of mode index.

Chapter 6

6 The Coupled Cluster Entropy estimation method for application to macromolecules

6.1 Introduction

The previous chapter established that non-parametric estimation of the configurational probability density is indeed possible in up to 45 dimensions. Within the Minimally Coupled Subspace (MCSA) framework, Full Correlation Analysis (FCA) is utilised to find coordinate transformations which provides a set of non-correlated eigenmodes, and subsequently almost non-correlated subspaces are identified. On each of these subspaces, the density estimation method introduced in the previous chapter is applied and an improved density estimate with respect to the quasi-harmonic approximation is obtained. The downside of this approach, however, is that FCA is computationally expensive and rests on linear coordinate transformations only. In case of slowly converging FCA and if linear transformations are inapt to find uncorrelated modes, sufficiently small independent subspaces to render non-parametric density estimation possible cannot be obtained.

To address this issue, in this chapter a method is introduced which shows how to split the configurational space (or oversized subspaces provided by the FCA) of macromolecules into arbitrary subspaces without neglecting correlations between those.

To this end, a correlation expansion is used which has hitherto been employed for the calculation of real gas properties [148] and since then is sometimes also referred to as the Kirkwood superposition approximation or mutual information expansion. Even though many attempts have been made so far [134, 136, 137] it has never successfully applied to

larger molecules of biological interest.

6.2 Theory

6.2.1 Inclusion-exclusion principle – Review

The MIE we are going to exploit here bears close resemblance to the principle of inclusion and exclusion known from set theory since the late 17th century. The inclusion-exclusion principle states that if $A_1, \ldots, A_n$ are finite sets one can estimate the cardinality $|\bigcup A|$ of the union of all sets as

$$\begin{aligned}\left|\bigcup_{i=1}^{n} A_i\right| &= \sum_{i=1}^{n} |A_i| - \sum_{i,j:1\leq i<j\leq n} |A_i \cap A_j| + \ldots \\ &+ \sum_{i,j,k:1\leq i<j<k\leq n} |A_i \cap A_j \cap A_k| - \ldots \\ &+ (-1)^{n-1} |A_1 \cap \cdots \cap A_n| \end{aligned} \tag{6.1}$$

If all sets $A_1, \ldots, A_n$ are disjoint then clearly $|\bigcup_{i=1}^{n} A_i| = \sum_{i=1}^{n} |A_i|$. The complex shape of this equation (as well as the name) results from alternating over-generous subtraction followed by addition of correction terms.

Application to Information Theory

Due to the additive property of measures one can directly transform this formula, which has been derived from elementary set theory, to probabilties, as has been shown by Poincaré [149]. In fact, Yeung proved that any set theoretical entity/operation can be viewed as information theoretical entity/operation [150, 151]. It is therefore not surprising [152] that the inclusion-exclusion principle can be shown to have an information theoretical twin expressing the cardinality of a set A as entropy, the intersect of two sets $(A_i \cap A_j)$ as Kullback-Leibler divergence or mutual information and the intersect of more than two sets $(A_i \cap \cdots \cap A_k)$ as interaction information or "mutual information of higher order". Thus, one can expand the entropy $S(A)$ of an arbitrary joint ensemble $A = \{A_1, \ldots, A_n\}$ in terms of lower order probabilities,

$$S(A_1,\ldots,A_n) = \sum_{i=1}^{n} S(A_i) - \sum_{i<j} I_2(A_i,A_j) + \ldots \\ \sum_{i<j<k} I_3(A_i,A_j,A_k) - \cdots + (-1)^{n-1} I_n(A_1,\ldots,A_n)\,, \quad (6.2)$$

with

$$I_n(A_i,\ldots,A_{i+n-1}) = \int_{i,\ldots,i+n-1} p_n(A_i,\ldots,A_{i+n-1}) \ln \frac{p_n(A_i,\ldots,A_{i+n-1})}{\prod_{j=1}^{n} p_1(A_j)}\,. \quad (6.3)$$

The rather complicated structure of these two intertwined equations is illustrated in Fig. (6.1), the information theoretical analogon of set theory's Venn diagrams, where circles are used to symbolize ensembles A, B and C and circle overlaps denote pair/third order correlations I_n. More detailed information is given by Yeung [150].

Superposition approximations

The nth order correlation functions

$$\tilde{p}_n(A_i,\ldots,A_{i+n-1}) = \frac{p_n(A_i,\ldots,A_{i+n-1})}{\prod_{j=1}^{n} p_j\,(A_j)}$$

can now be formally correctly expanded as a product of lower-order probabilities and a correction factor $\Delta^{(n)}$ [135]

$$\begin{aligned} \tilde{p}_n(A_i,\ldots,A_{i+n-1}) &= \frac{p_n(A_i,\ldots,A_{i+n-1})}{\prod_{j=1}^{n} p_j\,(A_j)} \\ &= \Delta^{(n)} \prod_{s=2}^{n-1} \left[\prod_{\{i^s\}} \tilde{p}_s(A_1,\ldots,A_s) \right]^{(-1)^{s+n-1}}, \end{aligned} \quad (6.4)$$

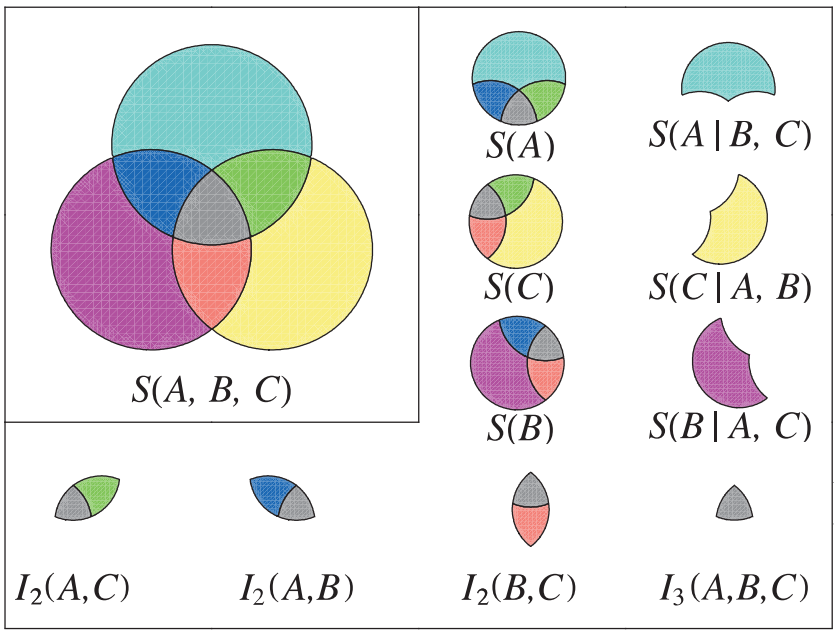

Figure 6.1: I-diagram for three ensembes A, B and C. Circles $S(A), S(B)$ and $S(C)$ denote the entropy of the these ensembles, the overlap between two symbolizes the pair-correlations $I_2(A,B), I_2(A,C)$ and $I_2(B,C)$, respectively. The overlap between all three ensembles stands for the three-body correlation $I_3(A,B,C)$, which in this representation cannot become negative (cf. Sec. 6.2.6). The non-overlapping parts are the conditional entropies $S(A|B,C), S(B|A,C)$, and $S(C|A,B)$ which become important for negative higher-order correlations.

where the inner product runs over all $\binom{n}{s}$ possible combinations $\{i_1, \ldots, i_s\} \in \{1, \ldots, n\}$. The remainders $\Delta^{(n)}$ are what keep the expansion formally exact. An approximation for the entropy can, however, be made by truncating the expansion at some point $t \leq n$. Note that truncating at a given order does not imply ignoring correlations beyond that order but making a particular superposition approximation for them, namely the one that has corrections of that order *and greater* set to unity. This has the important consequence that also Eq. 6.2 is truncated.

A well-known superposition approximation was suggested by Kirkwood. It formulates three-body correlation functions as the product of their three pairwise probabilities according to Eq. 6.4,

$$\begin{aligned} \tilde{p_3}(A_i, A_j, A_k) &= \tilde{p}_2(A_i, A_j)\tilde{p}_2(A_i, A_k)\tilde{p}_2(A_j, A_k) \\ &\quad \Delta^{(3)}(A_{i,}A_j, A_k)\,, \end{aligned}$$

which recovers Eq. 6.4, and is therefore exact, and then sets $\Delta^{(3)}(A_{i,}A_j, A_k) = 1$:

$$\tilde{p_3}(A_i, A_j, A_k) \approx \tilde{p}_2(A_i, A_j)\tilde{p}_2(A_i, A_k)\tilde{p}_2(A_j, A_k) \tag{6.5}$$

Even though Kirkwood introduced this only for three-body correlations [148, 153], the general approach of truncating the series of Eq. 6.4 at some point $t \leq n$ setting $\Delta^{(t)} = 1$ happens to be called generalized Kirkwood approximation [135]. For instance, for the four-body correlation functions we have

$$\begin{aligned} \tilde{p_4}(A_i, A_j, A_k, A_l) &= \frac{\tilde{p_3}(A_i, A_j, A_k)\tilde{p_3}(A_i, A_j, A_l)\tilde{p_3}(A_i, A_k, A_l)\tilde{p_3}(A_j, A_k, A_l)}{\tilde{p}_2(A_i, A_j)\tilde{p}_2(A_i, A_k)\tilde{p}_2(A_i, A_l)\tilde{p}_2(A_j, A_k)\tilde{p}_2(A_j, A_l)\tilde{p}_2(A_k, A_l)} \\ &\quad \Delta^{(4)}(A_i, A_j, A_k, A_l)\,. \end{aligned} \tag{6.6}$$

Setting $\Delta^{(4)} = 1$ yields the Fisher-Kopeliovich superposition approximation [154]. For comparison, applying the Kirkwood superposition approximation to this four-body problem yields

$$\begin{aligned} \tilde{p_4}(A_i, A_j, A_k, A_l) &= \tilde{p}_2(A_i, A_j)\tilde{p}_2(A_i, A_k)\tilde{p}_2(A_i, A_l)\tilde{p}_2(A_j, A_k)\tilde{p}_2(A_j, A_l)\tilde{p}_2(A_k, A_l) \\ &\quad \Delta^{(3)}(A_{i,}A_j, A_k)\Delta^{(3)}(A_{i,}A_j, A_l)\Delta^{(3)}(A_{i,}A_k, A_l) \\ &\quad \Delta^{(3)}(A_{j,}A_k, A_l)\Delta^{(4)}(A_i, A_j, A_k, A_l)\,, \end{aligned} \tag{6.7}$$

and setting $\Delta^{(4)} = 1$ and all $\Delta^{(3)} = 1$. This shows that indeed truncating Eq. 6.4 at $t \leq n$ also truncates Eq. 6.2.

Setting the correction factors $\Delta = 1$ in Eqs. 6.5, 6.6 or 6.7 bears the problem that then their left-hand side is normalised, whereas the right one is not, since for instance

$$\tilde{p}_2(A_i, A_j)\tilde{p}_2(A_i, A_k)\tilde{p}_2(A_j, A_k) = \frac{p_2(A_i, A_j)p_2(A_i, A_k)p_2(A_j, A_k)}{\left(p(A_i)p(A_j)p(A_k)\right)^2}\,.$$

Watanabe claimed that for this reason expressions like these are in general not meaningful and do not allow any useful interpretation [155]. Yet they provide the correct asymptotic behaviour in the case of vanishing correlations, i.e. as $I_2(A_i, A_k), I_2(A_i, A_k), I_3(A_i, A_j A_k) \rightarrow 0$ only the correlation between i and j remains. However, for higher correlations, the Kirkwood approximation greatly overestimates the triplet probability and higher-order analogues [156]. The validity of the Kirkwood approximation is therefore restricted to weak correlations. Correlations larger than $0.3 - 0.5$ have previously been found to introduce non-convergent behaviour as the third term overcorrects errors of the first two terms [135]. Hence, even though the generalised Kirkwood approximation holds the promise of constituting a suitable framework for obtaining configurational entropies in the high-dimensional configurational space of macromolecules, expansions over highly correlated degrees of freedom (e.g., internal coordinates) are likely to be problematic. To avoid divergencies of the expansion terms, weakly correlated collective coordinates, obtained from coordinate transformations as discussed in the previous chapters, are a suitable basis.

6.2.2 Application to PCA/FCA modes

Analogously to Eq. 6.2, the entropy $S\left[\rho\left(\mathbf{y}\right)\right]$ of an atomistic ensembles $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ of n configurations $\mathbf{x}_i \in \mathbb{R}^{3N}$ of a macromolecule of N atoms, after some coordinate

transformation $\mathbf{y} = \mathbf{T}\mathbf{x}$, can be expanded as

$$\begin{aligned} S\left[\rho\left(\mathbf{y}\right)\right] &= \sum_{i=1}^{3N} S\left[\rho\left(y_i\right)\right] - S_{\mathrm{res}}\left[\rho\left(\mathbf{y}\right)\right] \\ &= \sum_{i=1}^{3N} S\left[\rho\left(y_i\right)\right] - \sum_{i<j} I_2\left[\rho\left(y_i, y_j\right)\right] + \\ &\quad \sum_{i<j<k} I_3\left[\rho\left(y_i, y_j, y_k\right)\right] - \cdots + \left(-1\right)^{3N+1} I_{3N}\left[\rho\left(y_1, \ldots, y_{3N}\right)\right] , \end{aligned} \tag{6.8}$$

where y_i denote the components of $\mathbf{y} = \mathbf{T}\mathbf{x}$.

The most naive and straightforward approach would be to estimate the marginal distributions $\rho\left(y_i\right) = \int \rho(\mathbf{y}) dy_{j\neq i}$ from a quasi-harmonic (QH) approximation, which, as discussed in the previous chapters and in Chapter 2, is obtained by fitting the configurational density $\rho(\mathbf{x})$ with a multi-variate gaussian function [56, 57]. The configurational density approximately factorizes,

$$\rho_{\mathrm{QH}}\left(\mathbf{y}\right) = \prod_{i=1}^{3N} \rho\left(y_i\right), \tag{6.9}$$

but here only linear correlations are eliminated. Non-linear and higher-order correlations, in contrast, prevail and consequently the entropy

$$S\left[\rho\left(\mathbf{y}\right)\right] = \sum_{i=1}^{3N} S_{\mathrm{QH}}\left[\rho\left(y_j\right)\right] + S_{\mathrm{res}}\left[\rho\left(\mathbf{y}\right)\right]$$

contains a non-neglegible residual entropy $S_{\mathrm{res}}\ \left[\rho\left(\mathbf{y}\right)\right]$ due to ignored correlations. Approximating $S_{\mathrm{res}}\ \left[\rho\left(\mathbf{y}\right)\right]$ according to Eq. 6.8 could, in principle, substantially improve the QH result. From a computational point of view one has now replaced the evaluation of the $3N$ dimensional integral, $\int_{3N} \rho \ln \rho$, by the evaluation of $\sum_{i=1}^{t} \binom{3N}{i}$ i-dimensional probability density estimations, t being the truncation order for which a particular superposition approximation is being made. Already for fairly low $\{N, t\}$ one is confronted with an unsurmountable combinatorial explosion. Attempts to apply this scheme to biological macromolecules have failed due to the high number of terms to be evaluated [157**?**].

6.2.3 Mode clusters

To reduce the combinatorial pressure, we propose to make use of the fact that the entities whose correlations are sought can be completely arbitrarily chosen by using groups of modes rather than individual modes.

Accordingly, given disjoint sets of eigenmodes C_s we cluster into h_s disjoint subclusters $D_i^{(s)}$ of sizes $d_1^s, \ldots, d_{h_s}^s < 15$, so that $C_s = \bigcup_{i=1}^{h_s} D_i^{(s)}$. C_s may either be the complete set of $3N$ PCA or FCA modes or, within the Independent Subspace framework, oversized clusters. The approximation to the residual entropy then reads

$$S_{\mathrm{res}}\left[\rho\left(C_s\right)\right] = -\sum_{i<j} I_2^{(d_i^s+d_j^s)}\left[\rho_i,\rho_j\right] + \sum_{i<j<k} I_3^{(d_i^s+d_j^s+d_k^s)}\left[\rho_i,\rho_j,\rho_k\right] - \cdots + (-1)^{h_s+1} I_{h_s}\left[\rho_i,\ldots,\rho_{h_s}\right] , \quad (6.10)$$

where $\rho_i \equiv \rho^{(d_i^s)}\left(\bigotimes_{j\in D_i^{(s)}} y_j\right)$. This has three advantages. First, it significantly retards the combinatorial explosion since now one has to evaluate only $\sum_{i=1}^{t}\binom{3N/d_i^s}{i}$ correlation terms. Furthermore, inside every group one has effectively taken *all* correlations up to order d_i^s into account without making any superposition assumption at all. That is, only the d_i^s -th fraction of the original numbers of correction factors $\Delta^{(n)}$ will be set to unity resulting in less severe ignorance about the true behaviour of the system. This in turn will, third, result in faster convergence of the expansion, since less terms in Eq. 6.4 need to be overcorrected for with higher-order correlations.

6.2.4 Error cancellation via 'fill modes'

We applied the non-parametric density estimation method developed in the last chapter to obtain configurational probability density estimates $\rho\left(y_i\right)$. The configurational entropy is estimated according to

$$S_c = \frac{k_{\mathrm{B}}}{n}\sum_{i=1}^{n} \ln \frac{n Z_d\left(\mathbf{y}_i\right) r_{i,k}^d}{k\left(\mathbf{y}_i, r_{i,k}\right)} , \quad (6.11)$$

where $k\left(\mathbf{y}_i, r_{i,k}\right) = \left\langle K\left(\mathbf{y}_i, \left(\mathbf{y}_i - \mathbf{y}\right)/r_{i,k}\right)\right\rangle_{\mathbf{y}}$ denotes the ensemble average of a locally adapted kernel function K, whose anisotropy and scaling $r_{i,k}$ depends on the local density at point $\mathbf{y}_i$. Due to the moderate regularization assumptions, this estimator is sensitive to the sparse sampling problem whose effect is highly dependent on the dimensionality

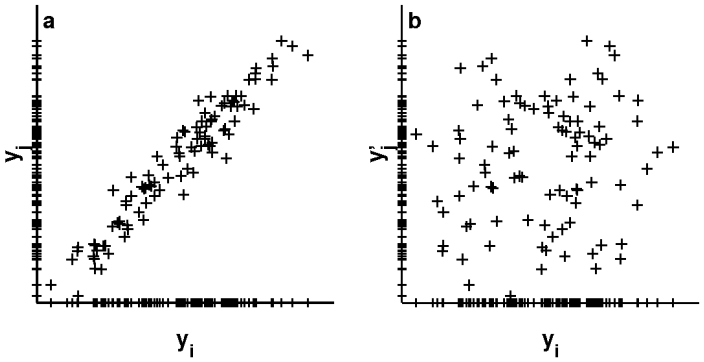

Figure 6.2: a) Two arbitrary strongly correlated modes y_i and y_j marginally distributed on the axes. Correlation is clearly visible from the y_j-distributed y_i. The joint distribution $p(y_i, y_j)$ is more sparsely sampled than both marginal distributions. b) The y'_j-distributed y_i is decorrelated and has exactly as many sample points as the joint distribution in a) allowing precise computation of $I_2(y_i, y_j)$.

(see also Sec. 5.3.2). To guarantee the same accuracy of all density estimates required for the computation of the correlation terms I_n of Eq. 6.10 despite different dimensionality, it is necessary to ensure the same local densities around points $\mathbf{y}_i$ in different terms. This is normally not provided. The mutual information between two modes y_i and y_j,

$$I_2 = \int_{i,j} \rho(y_i, y_j) \ln \frac{\rho(y_i, y_j)}{\rho(y_i)\rho(y_j)} \,. \tag{6.12}$$

contains differently well sampled terms in denominator and numerator, because the number of sampling points available to estimate $\rho(y_i, y_j)$ is only half the number of sampling points available for estimating the marginal densities $\rho(y_i)$ and $\rho(y_j)$ (cf. Fig. 6.2). The accuracy for the estimation of the marginal densities is, consequently, possibly higher than the joint estimate yielding an inaccurate correlation estimate. To overcome this problem, we devised the concept of fill modes. Accordingly, artificially decorrelated modes $y'_i : \{y'_{i,1}, \ldots, y'_{i,3N}\} = \mathrm{perm}\{y_{i,1}, \ldots, y_{i,3N}\}$ are created by randomly permuting its components $y_{i,j}$, with $1 \leq j \leq 3N$. The marginal densities $\rho(y'_i) = \rho(y_i)$ and $\rho(y'_i, y_j) = \rho(y_i)\rho(y_j)$, yielding a new expression for Eq. 6.12,

$$I_2 = \int_{i,j} \rho(y_i, y_j) \ln \frac{\rho(y_i, y_j)}{\rho(y'_i, y_j)} \,, \tag{6.13}$$

where the product of the marginal densities $\rho(y_i)$ and $\rho(y_j)$ is now computed from the synthetically decorrelated joint distribution $\rho(y_i', y_j)$, such that the same accuracy for the joint estimate is guaranteed as for the marginal estimates. Conducting this scheme on the third order correlation function of three modes y_i,y_j and y_k,

$$I_3 = \int_{i<j<k} \rho(y_i, y_j, y_k) \ln \frac{\rho(y_i, y_j, y_k)}{\frac{\rho(y_i,y_j)\rho(y_i,y_k)\rho(y_j,y_k)}{\rho(y_i)\rho(y_j)\rho(y_k)}} ,$$

yields

$$I_3 = \int_{i<j<k} \rho(y_i, y_j, y_k) \ln \frac{\rho(y_i, y_j, y_k)}{\frac{\rho(y_i,y_j,y_k')\rho(y_i,y_k,y_j')\rho(y_j,y_k,y_i')}{\rho(y_i',y_j',y_k')^2}} , \tag{6.14}$$

where the pairwise joint distributions have been 'filled up' with randomly permuted 'fill modes', as described above, e.g. $\rho(y_i, y_j) = \rho(y_i, y_j, y_k')/\rho(y_k')$. The same scheme has also been applied to correlation functions of higher order than three (c.f. Fig. 6.5).

6.2.5 Consistent dimensions

The sensitivity of the nearest-neigbour estimates, Eq. 6.11, towards the sparse sampling problem also affects the different terms of Eq. 6.10, which inevitably suffer from different sparse sampling problems if computed seperately. Furthermore, a huge number of probability density distributions $\rho(y_i), \rho(y_i, y_j), \ldots, \rho(y_i, y_j, \ldots, y_k)$ is computed more than once for the many instances of identical correlation terms appearing in that equation. Expanding over entropy terms rather than correlation terms, in contrast, yields

$$S\left[\rho\left(y_1, \ldots, y_n\right)\right] = \sum_{k=1}^{t} \sum_{m_1<\ldots<m_k} \theta_{k,t} S\left[\rho\left(y_1, \ldots, y_k\right)\right], \tag{6.15}$$

where the first summation runs over different orders $k = 1, \ldots, t$ until truncation order $t \leq n$. $\theta_{k,t} = \sum_{i=k}^{t} (-1)^{i+k} \binom{n-k}{i-k}$ designates how many times a certain order appears and whether it needs to be added or subtracted, and the second sum over all $\binom{n}{k}$ possible combinations $\{m_1, \ldots, m_k\} \in \{1, \ldots, n\}$. Table 6.1 summarizes a few features of this expansion. In A, the θ_k, $k = 1, \ldots, 4$ have been calculated for 4 arbitrary modes y_i. As can be seen, for $t = n$ all lower-order terms vanish and only the highest-order term remains with a weighting factor of exactly 1. B–C show the θ_k for different numbers of modes $n = 10, 20, 100$ when expanding up to order 4. As can be seen, for fairly low n,

A $n = 4$

k/t	1	2	3	4
1	1	-2	1	0
2	0	1	-1	0
3	0	0	1	0
4	0	0	0	1

B $n = 10$

k/t	1	2	3	4
1	1	-8	28	-56
2	0	1	-7	21
3	0	0	1	-6
4	0	0	0	1

C $n = 20$

k/t	1	2	3	4
1	1	-18	153	-816
2	0	1	-17	136
3	0	0	1	-16
4	0	0	0	1

D $n = 100$

k/t	1	2	3	4
1	1	-98	4753	-152096
2	0	1	-97	4656
3	0	0	1	-96
4	0	0	0	1

Table 6.1: Factors θ_k of Eq. (6.15) as a function of truncation orders t and index k, for four different mode/subcluster numbers n.

these weighting factors grow very large. This illustrates how much computational time can be saved when expanding the entropy in terms of Eq. 6.15 rather than Eq. 6.8.

To guarantee the same estimation accuracy for all $\rho(y_1, \ldots, y_k)$ of Eq. 6.15, each term is filled up to truncation order t yielding $\rho\left(y_1, \ldots, y_k, y'_{k+1}, \ldots, y'_t\right)$. Under this modification, Eq. 6.15 reads

$$\begin{aligned} S\left[\rho\left(y_1, \ldots, y_n\right)\right] &= \theta'_{1,t} \underbrace{\sum_{m_1,\ldots,m_t} S\left[\rho\left(y'_1, \ldots, y'_t\right)\right]}_{\text{marginal entropies/fill modes}} + \\ &\sum_{k=2}^{t} \sum_{m_1<\ldots<m_k} \theta_{k,t} S\left[\rho\left(y_1, \ldots, y_k, y'_{k+1}, \ldots, y'_t\right)\right] , \end{aligned} \tag{6.16}$$

with the number of marginal entropies,

$$\theta'_{1,t} = \underbrace{\sum_{i=1}^{t} (-1)^{i+1} \binom{n-1}{i-1}}_{\text{normal first-order indexing}} - \underbrace{\sum_{i=2}^{t-1} \theta'_t \frac{\binom{n}{i}\binom{n-i}{t-i}(t-i)}{n}}_{\text{fill modes}} ,$$

which depends on the fill mode weighting index

$$\theta'_t = \sum_{k=2}^{t} \sum_{i=k}^{t} (-1)^{i+k} \binom{n-k}{i-k},$$

where, like above, primes indicate permuted entries.

6.2.6 Negative Correlations

Higher-order correlation functions can, in principle, be either positive or negative. Systems with negative higher-order correlations are called frustrated [158] and complex macromolecules are in many respects paradigms for frustrated systems [159]. It is sometimes argued [157] that one can therefore expect higher-order correlations of proteins always to be negative. The three-body correlation of three modes y_i, y_2, y_3 can be rewritten as

$$I_3(y_i, y_j, y_k) = I_2(y_i, y_j) + I_2(y_i, y_j) - I_2(y_i, y_j y_k)$$

where

$$I_2(y_i, y_j y_k) = S\left[\rho\left(y_i\right)\right] + S\left[\rho\left(y_i, y_k\right)\right] - S\left[\rho\left(y_i, y_j, y_k\right)\right]$$

is the information gained of y_i from simultaneous measurement of y_j and y_k. Since none of the terms appearing in the former equation goes into the latter, it is possible that $I_2(y_i, y_j y_k)$ is nonzero even if $I_2(y_i, y_j)$ and $I_2(y_i, y_k)$ (nearly) vanish. $I_3(y_i, y_j, y_k)$ will be negative if the information gained of y_i by simultaneous measurement of y_j and y_k is larger than the information gained of y_i from separate measurement of y_j and y_k [158].

To illustrate this concept, Fig. 6.3 rather sketchily displays an arbitrary rugged object (a) which has the property that neither of the three 2D-projections (b–c) conveys the information needed to give the true shape of this object. All 2D projections suggest a cubic shape, but the total volume of the object depicted in Fig. 6.3 is clearly much smaller than that of a cube of the same edge length, which can be verified by assuming a 3D view corresponding to including third-order information into the observation. Hence, the 'view' on the high-dimensional landscape is what determines how many modes are needed for assessing the true extent of it. This is independent of the systems' frustrated properties since proper choice of $\mathbf{T}$ can always return a set of uncorrelated modes $\mathbf{y}$. Frustration, thus, does not describe the inherent properties of the system but rather the quality of the applied coordinate transformation $\mathbf{T}$. This fact will become important for

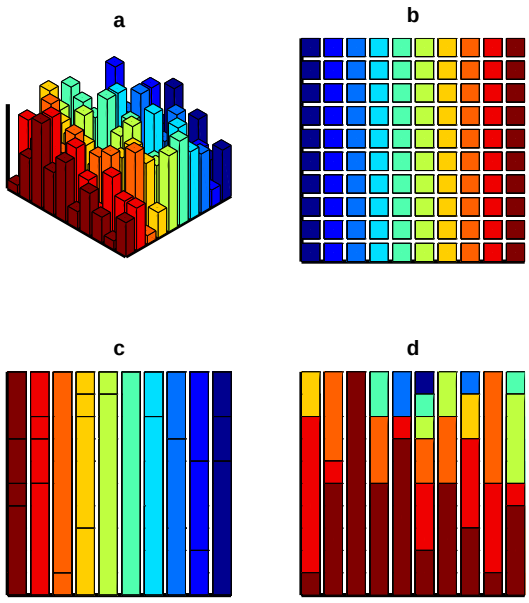

Figure 6.3: a) An arbitrary rugged column object with negative third-order correlation I_3. Colours are for better orientation only. Neither of the three 2D projections (b–d) conveys the information needed to assess the true extent of the object, which from the 2D views appears to be a cube. Including third-order information reveals that the true volume is in fact much smaller than a cube's volume of given edge length.

the interpretation of expansion curves of alkanes in Sec. 6.3.2.

6.3 Results

To test the applicability and accuracy of the above discussed expansions, we examined a number small alkane systems, hexane, octane and decane. As shown in the previous chapter, due to their flexibility, these constitute particularly hard test examples. It can be expected, that also here they are suitable to explore the limits of the introduced method. Of particular interest are convergence properties. The generalized Kirkwood approximation has been shown to diverge in case of highly correlated systems (Sec. 6.2.1). Furthermore, it is a priori unclear whether correlation functions of order > 2 are positive or negative. It is of high interest to see if these expansions still yield upper bounds to the true entropy.

6.3.1 Single mode vs. clustered mode expansions

The six-atomic alkane hexane modelled with an united-atom force field was examined as a first test example, a 18 dimensional estimation problem (Fig. 6.4). A reference value from a thermodynamic integration (TI) scheme (dashed line) was obtained, as described in the previous chapter. A further reference value is the direct density estimate, obtained via conducting locally adapted anisotropic kernel density estimation in the full 18-dimensional configurational space (dotted line). This value is slightly larger than the thermodynamic reference (for exact numbers cf. Table 5.1). For comparison, the QH estimate, which ignores all non-linear correlation between the 18 eigenmodes, is shown as a solid line.

In Fig. 6.4, the square-mark line shows the expansion of individual PCA modes according to Eq. 6.8. Four of the 18 PCA eigenmodes were highly correlated with correlation factors above 0.3. Subtraction of all $\binom{18}{2} = 153$ possible pair-correlations, which were calculated according to Eq. 6.13, gave rise to a value well above the reference. Including all $\binom{18}{3} = 816$ third order terms, too, which were computed according to Eq. 6.14, showed dramatically diverging behaviour underestimating the configurational entropy by a wide margin (Note that the y-axis is broken to accomodate the whole data range spanned by this plot). The large underestimation occurs even though consistent sampling as discussed in Sec. 6.2.4 was ensured for both I_2 and I_3. Without fill modes, the underestimation was even larger (data not shown).

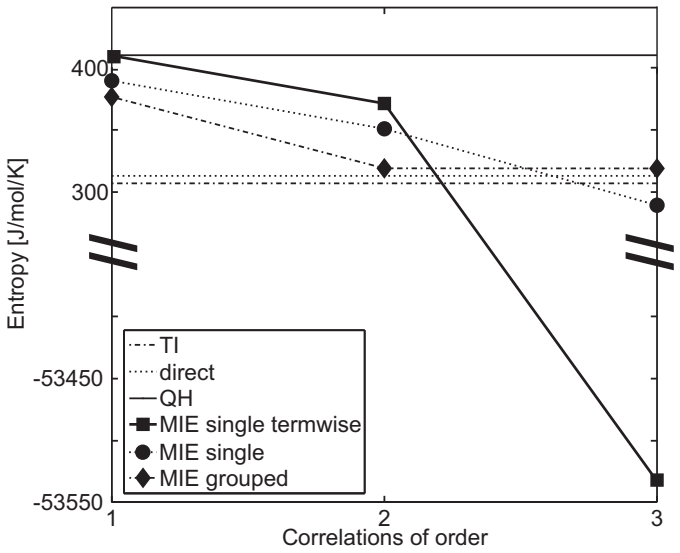

Figure 6.4: Configurational entropy of hexane estimated with different approaches. Dashed line: TI reference entropy; Dotted line: Locally adapted kernel direct density estimate on the whole configurational space; Solid line: QH density estimate; Solid lines with squares: Computation of Eq. 6.2 with single modes (no clustering) with sum of marginal entropies (first term) from the quasi-harmonic appoximation, the second-order correlations from Eq. 6.13 and the third-order correlations from Eq. 6.14; dotted line with circle: computation of all three terms based on Eq. 6.16 with single modes (no clustering); Dashed line with diamonds: computation of all three terms based on Eq. 6.16 with configurational space split into six clusters.

The same data as before were used for the computation of Eq. 6.16 ensuring consistent sampling not only within each correlation term but throughout all terms up to truncation order $t = 3$ as described before. The resulting plot "MIE single" in Fig. 6.4 clearly shows that most of the previous underestimation can be attributed to the bias introduced by inconsistent sampling densities of different order correlation functions. Nevertheless, a non-negligible underestimation by $\sim 30\,\mathrm{JK}^{-1}\mathrm{mol}^{-1}$ ($\sim 10\%$) remains. Hence, these types of expansions on single, potentially highly correlated PCA-modes do not yield upper bounds to the true entropy as, for example, Schlitter's quasi-harmonic approximation [57] or the adaptive kernel density estimations. Furthermore, the combinatorial pressure of expanding up to order three is quite remarkable for a molecule of this size. In total, the computation of Eq. 6.16 for 18 PCA modes up to order three required 987 single

computations.

"MIE grouped" in Fig. 6.4 shows the convergence behaviour of expansions on the basis of mode clusters rather than single modes. Here, the 18 modes were split into 6 groups of 3 modes each, and only 41 terms needed to be computed and used for evaluation of Eq. 6.16. Clearly, the expansion has converged after the second term leaving a deviation of only $\sim 10\,\mathrm{JK}^{-1}\mathrm{mol}^{-1}$ to the direct density estimate which includes all relevant correlations of the system. Furthermore, the clustered expansion did return an upper bound to the true entropy.

The clustered expansion yields more accurate results at lower computational cost. This could be a property of hexane rather than a generally valid observation. In the following sections, the convergence properties of clustered expansions as a function of cluster number will therefore be explored.

6.3.2 Influence of the number of clusters on convergence

Octane and decane comprising 8 and 10 atoms, respectively, served as our next test systems. The molecules were simulated with the united atoms force-field ffG43b1 [65], and an ensemble of 500.000 sample points from a stochastic dynamics simulation was used to estimate the configurational entropy. Full Correlation Analysis (FCA) was carried out to obtain a set of optimally decoupled eigenmodes.

Fig. 6.5 shows the obtained MIE estimates for octane (top row) and decane (bottom row). The TI reference entropies are shown as solid horizontal lines. Also given are, for comparison, the density estimates obtained by locally adapted anisotropic kernel density estimation of the whole configurational space, S_{dir} (dashed-dotted horizontal lines), and estimates according to the minimally coupled subspace framework (MCSA), S_{MCSA} (dotted horizontal lines) using a clustering threshold of $\theta = 0.025$ as described in the previous chapter. For both octane and decane, the expansions were calculated for using Eq. 6.16 choosing truncation orders $t = 3$ (left column) and $t = 4$ (right column). In each case, the clustering threshold θ was chosen such that uniformly sized mode clusters could be formed. Remaining modes not assigned to an expansion cluster were subjected to separate adaptive kernel density estimation, and the entropy estimate was added to the expansion estimate at each permutation order.

For octane (top row), the expansions were conducted with four (circles), five (squares), six (diamonds) and eight (triangles) mode clusters, respectively. For truncation at order

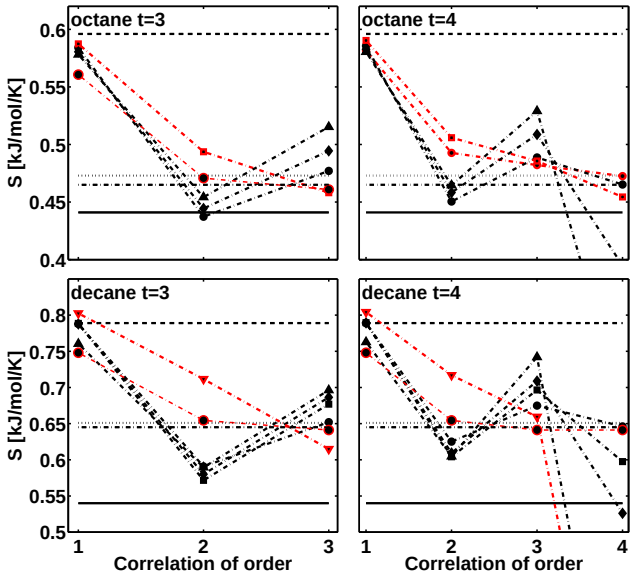

Figure 6.5: Octane (top) and decane (bottom) expansions calculated at truncation orders $t = 3$ (left) and $t = 4$ (right). The configurational space was split into 4 (circles), 5 (squares), 6 (diamonds), 8 (upwards triangles, octane), 9 (upwards triangles, decane), and 12 (downwards triangles, decane $t = 4$ only) clusters, respectively. Red: Expansions on independent subspace; black: expansions on total configurational space. Solid horizontal lines: TI reference entropies; dashed-dotted horizontal lines: direct adaptive kernel density estimates on the whole configurational space; dotted horizontal lines: independent subspace estimates.

$t = 3$ (upper left), certain expansions exhibit zig-zagging convergence behaviour (shown in black), whereas others converge more monotonously (in red). Furthermore, increasing cluster count seems to delay the onset of convergence.

The picture is similar for truncation at order $t = 4$ (upper right). Here, too, certain expansion are zig-zags whereas others are not. Like before, increasing cluster count entails more pronounced amplitudes indicating delayed convergence. The third-order values obtained here are only slightly larger than those from the previous plot. The fourth-order values, in contrast, diverge with increasing cluster count. On the one hand, as required, the four-cluster plot (circles) reproduces the direct density estimate (dashed-dotted) at permutation order 4, where, according to Eq. 6.16, all the lower-order contributions

vanish. The five-cluster expansion, on the other hand, yields an estimate below the direct density estimate but still above the TI reference, and the expansions with higher cluster counts even yield values well below the TI reference.

For decane (lower row), a similar situation is seen. Here, the expansions were carried out with four (circles), five (squares), six (diamonds), nine (upward triangles) and twelve (downward triangles) mode clusters, respectively. Like before, the four-cluster expansions recover the MCSA estimate for $t = 3$ (left-hand side), and the direct density estimate for $t = 4$ (right-hand side), as required. Increasing cluster counts also delays convergence as seen from increasing amplitude; and the fourth-order permutation does not yield sensible values for more than five clusters.

In summary, Fig. 6.5 clearly shows that correlation expansions can improve on the QH approximation. Notably, and in contrast to the diverging fourth-order estimates, neither second-order nor third-order estimates underestimate the TI reference. Particularily the second-order estimates are very close to the TI-reference. Is it therefore justified to truncate the expansions after the second term? It is problematic that the second-order estimates yield values below the direct density estimates in almost all cases. As can be seen from the four-cluster expansions for both octane and decane, expansions need to converge to the direct density estimate since here all the lower-order terms cancel out. Thus, while it is, in principle, desired to recover the TI reference, the fact that the second-order estimates are so close appears to be due to accidental error cancellations of the expansions rather than the systems' property.

It is furthermore evident that the cluster count does play an important role for the accuracy of the entropy estimate. More clusters entail delayed convergence or even pronounced divergence, particularily at permutation order four, where significant underestimation of the TI reference is obtained if the number of clusters exceeds five. We attribute the divergence at permutation order four to sparse-sampling problems. The large number of terms to be evaluated at the fourth order with, e.g., nine or twelve clusters requires extremely accurate entropy estimates for each single term. Such high accuracy might not even be provided by the adaptive kernel estimation method, as seen from the convergence issues of the larger alkanes illustrated in the previous chapter.

Interestingly, all mononously decreasing expansions (in red) were conducted on a single rather homogeneous independent subspace created according to the ISA scheme rather than on (almost) the whole, heterogeneous, configurational space. Minimally coupled subspaces are created based on a paiwise correlation criterion; and as noted before, is is

assumed that pairwise uncorrelated modes do not exhibit correlations of higher-order either. It is furthermore assumed that the modes within a certain subspace consitute a network of correlated collective coordinates. The expansions of Fig. 6.5 corroborate this notion. Within the minimally coupled subspaces, pairwise correlations are relatively high (larger than 0.025, the clustering threshold). These pairwise correlations, however, still miss a considerable part of the multi-body correlation present in the network. Thus, the third-order correlation is negative (c.f. Sec. 6.2.6).

Expanding over correlations of the whole configurational space rather than over each subspace separately, the majority of correlation terms involve inter-cluster correlations; and these remaining low pairwise correlations seem to considerably overestimate the total inter-cluster correlations. This effect counterbalances the underestimation effect of the intra-cluster terms. Hence, the total third-order correlations are positive.

6.3.3 Application to Calmodulin

Next, we explored the characteristics of entropy expansion applying it to a macromolecule of biological interest, calmodulin. Modelled with the OPLS full atom force field in explicit solvent, it comprises a ~ 6500 dimensional optimization problem.

Using the entropy expansions for the purpose of improving on the quasi-harmonic result applying it to the whole configurational space would not be feasible since the number of single computations (i.e. the number of correlation terms) is $\sum_{i=1}^{3} \binom{6500}{i} = 4.6 \times 10^{10}$ when expanding on PCA mode pairs and triples, and still $\sum_{i=1}^{3} \binom{650}{i} = 4.6 \times 10^{7}$ when expanding on clusters of 10 thereof.

Following the MCSA scheme, Full Correlation Analysis was conducted on the first 2500 of the ~ 6500 PCA modes assuming quasi-harmonic characteristics for the remaining modes. After 3000 FCA optimization steps, the FCA modes were clustered as described above with two different thresholds, 0.22 and 0.03, respectively, yielding a maximum subspace size of 96 or 64, respectively. The entropy estimate based on the latter is expected to be considerably larger than the former, because more non-linear correlations are neglected. Second, from the alkane expansion, one would expect that lower cluster count yields lower amplitudes of the expansion plots. Third, different clustering should leave the entropy estimate at perturbation orders two and three roughly unchanged.

To verify these assumptions, we split the 96 modes subspace resulting from a clustering threshold of 0.22 into 9 or 12 groups, repectively, and the 64 subspace into $8, 9$ or 16

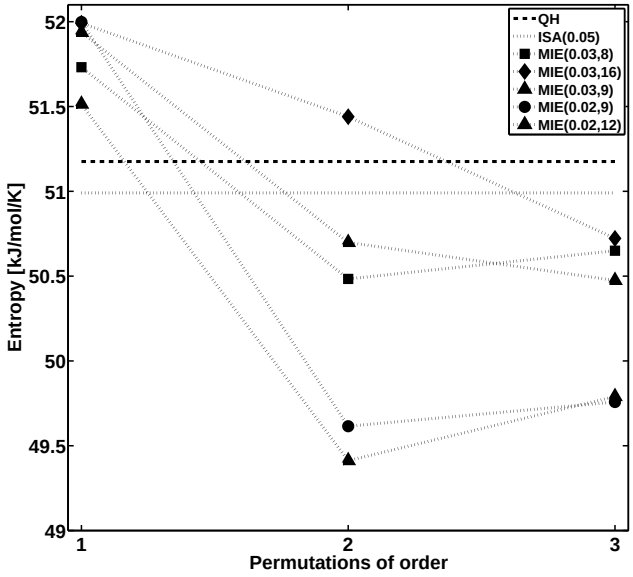

Figure 6.6: MIE for calmodulin with different cluster numbers and clustering thresholds. QH: quasi-harmonic estimate; FCA+density estim.: locally adapted anisotropic kernel density estimation on each of the minimally coupled subspaces obtained by clustering of the FCA modes using a high clustering threshold of 0.05; FCA+MIE (x, y): MIE conducted on each of the minimally coupled subspaces obtained by clustering with a clustering threshold of x, splitting the subspaces into y clusters.

groups. In Fig. 6.6, the two cases are plotted, together with two reference values; first, the QH approximation reference which was obtained by assuming quasi-harmonic behaviour for all CaM modes (dashed line), and the second reference is the estimate obtained by choosing a clustering threshold of 0.05, which results in subspaces small enough to be computed without any further entropy expansion (dotted line). While this second reference markedly improves on the quasi-harmonic estimate, the high threshold does not include sufficient non-linear correlations. Lower thresholds are, thus, likely to considerably improve on this estimate.

Indeed, a lower threshold of 0.03, which includes more correlations but requires MIE, yields considerably lower estimates in all three case and, importantly, the cluster number/size indeed leaves the entropy estimate almost unchanged. An even lower threshold of 0.022 further improves the estimate and the outcome is likewise the same regardless of whether the group size was 9 or 12. Interestingly, despite the size of the CaM configurational space the CaM expansion curves do not fluctuate as much as the alkane expansions. The third-order estimates do not deviate from the second-order estimate by more than 10%. This corroborates the previous observation that the alkanes constitute particularily hard test cases.

6.4 Conclusions

In this chapter, we showed that application of the set theoretical inclusion-exclusion principle markedly improves estimates of configurational entropies in configurational (sub-) spaces of macromolecules. We also showed that even the application to large biomolecules is possible if this method is embedded within the minimally coupled subspace approach (MCSA) framework developed in this thesis. The application of these mutual information expansions (MIE) is useful whereever the sparse sampling problem, the 'curse of dimensionality', impedes application of direct density estimation on the full minimally coupled subspace. Previous attempts to apply similar expansions to macromolecules failed due to the "combinatorial explosion" caused by the large number of correlation terms to be considered. We have shown that by clustering modes for the expansions improved results are obtained and the "combinatorial explosions" are alleviated. The problem of negative correlations has been analyzed and discussed. In particular, the appearance of negative higher-order correlations has been shown to be connected to 'frustrated' coordinate transformations rather than frustrated systems.

Outlook We here used generalised-Kirkwood type of correlation expansions and found that they are applicable in this framework. In principle, however, there is no reason to confine mutual information expansions to the generalised Kirkwood expansions. In information theory, Markov expansions have been shown to be more robust than Kirkwood expansions in the high-correlation regime. It would be interesting to test the applicability and usefulness for the estimation of configurational entropies.

Chapter 7

Allosteric regulation of pyruvate kinase

7.1 Introduction

Cell regulation relies vitally on enzymatic control at the molecular level both within and between molecules. Allostery is a mechanism to exert intramolecular control by modulating the activity of a protein's *catalytic* binding site by binding of a ligand at a distant *allosteric* binding site. Even though Pauling had proposed an allosteric model for the positive cooperativity of hemoglobin [160], the term *allosteric* was coined by Jaques Monod and François Jacob [161] more than 25 years later to describe such remote control mechanisms. The Pauling model was taken up in 1966 by Koshland [162] and is now generally referred to as the KNF model.

Modern understanding of allosteric mechanisms in proteins began with the symmetry, or MWC, model [42, 163], which sought to explain the Hill equation of binding oxygen to hemoglobin [164, 165] in terms of a structural difference between apo- and holo-hemoglobin [166, 167]. The symmetry model postulates, first, allosteric molecules to be made of identical monomers in a symmetric arrangement and, second, the allosteric effect to be due to a symmetry preserving rotation of the subunits upon binding of at least one ligand, which changes the quarternary structure from an inactive T-state (tensed) to an active R-state (relaxed). In this model, the different (T and R) quarternary structures are accessible in both the liganded and the unliganded form. However, binding of a ligand at the allosteric site of at least one monomer entails a shift of the T-R-equilibrium towards the R-state since the subunits are coupled in such a way that they always assume the same conformation (symmetry model). The symmetry/MWC model has found widespread

application in biology [42, 168] and has inspired a number of generalizations [169–175].

The Pauling-KNF or sequential model [160, 162], in contrast, drops the symmetry constraint and allows the subunits to assume different tertiary conformations. Moreover, instead of an T-R equilibrium (like in the symmetry model), the transition from T- to R-state of each subunit is assumed to be due to an induced fit mechanism. While such an induced fit converts a subunit from the tensed state to relaxed state, it does not propagate the conformational change to adjacent subunits. Instead, substrate-binding at one subunit only slightly alters the structure of other subunits so that their binding sites are more receptive to substrate. This allows for a larger number of intermediate states between 'full' R and 'full' T state.

The severe restrictions imposed on the behaviour of the subunits in either of these models have allowed for an elegant mathematical framework which found widespread appreciation in particular for the paradigm of allosteric proteins, hemoglobin. The applicability of either model or of one of their more sophisticated generalisations used to be, and still is, mainly tested by kinetic and thermodynamic experiments, in which the protein can be represented at the cartoon level and otherwise treated as a black box. Accordingly, the focus on the most pronounced structural changes seen in crystallographic studies of proteins, the quarternary change, sufficed. The kinetics of some allosteric systems seem to be correctly described by the symmetry model, others by the sequential model and some by neither model [176].

It is, however, problematic to relate kinetic measurements to a mechanical model [177]. Accordingly, it is now clear that almost all of these postulates do not apply to many systems studied so far. In particular, allosteric proteins have been discovered, which are only monomeric [44, 178, 179], the symmetry assumption is not always fulfilled [175], and quarternary changes are sometimes not observed in allosteric proteins [44]. Neutron scattering experiments found evidence for changes of protein dynamics upon ligand binding [180, 181] – data which corroborated newer models of allostery introduced previously by Weber [182], which emphasize the importance of continuous conformational distributions [183, 184]. Cooper and Dryden later realised [44] that ligand binding at an allosteric site can influence binding at a remote site without inducing mean conformational change, solely through alteration of atomic fluctuations. This finding inspired many follow-up studies of allosteric protein dynamics focusing mainly on effects of allosteric regulation on free energies [41, 185–187]. Moreover, since both MWC and Pauling-KNF model were primed for comparison with kinetic measurements, they are, by construction,

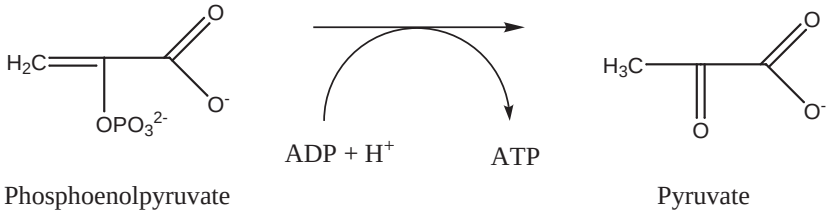

Figure 7.1: The reaction catalyzed by pyruvate kinase

phenomenological and as such do not answer at atomic detail the question of how the binding of a ligand result in an allosteric effect.

In this chapter, the allosteric regulation of pyruvate kinase is examined in atomic detail using extensive molecular dynamics simulations. This project was carried out in collaboration with the structural biology group of Malcolm Walkinshaw in Edinburgh, Scotland. Experimental data as well as some of the basic illustrations (marked as such) were provided by Malcolm's PhD student Lindsay Tulloch, who crystallised the R-state pyruvate kinase. Malcolm also suggested many simulations from an experimental point of view during the start-up phase of this project and my PhD time and acted as supervisor-in-charge for the first 1.5 years, when I was based in his lab. All work related to simulations and data analysis was carried out by myself.

7.2 L. mexicana PYK

Pyruvate kinase (PYK) catalyzes the final step in glycolysis, producing the second of two ATP molecules generated in the glycolytic pathway. In the presence of one equivalent of monovalent cations, normally K^+, and two equivalents of bivalent cations, usually Mg^{2+} or Mn^{2+}, the enzyme converts phosphoenolpyruvate (PEP) and ADP to pyruvate and ATP (Figure 7.1). PYK plays a central role in the cellular metabolism. The regulation of PYK is important for controlling the levels of ATP, GTP and glycolytic intermediates. PYK also serves as a switch between the glycolytic and gluconeogenic pathways in certain tissues [188]. PYK has been characterized from a number of prokaryotes and eukaryotes [189], and in nearly all organisms it shows allosteric properties in binding the substrate PEP. PYK from prokaryotes is normally activated homotropically by PEP and heterotropically by sugars bearing either one (i.e. ribose-phosphate) or two (i.e. fructose-1,6-biphosphate or fructose-2,6-biphosphate, respectively, FBP) phosphate groups. FBP

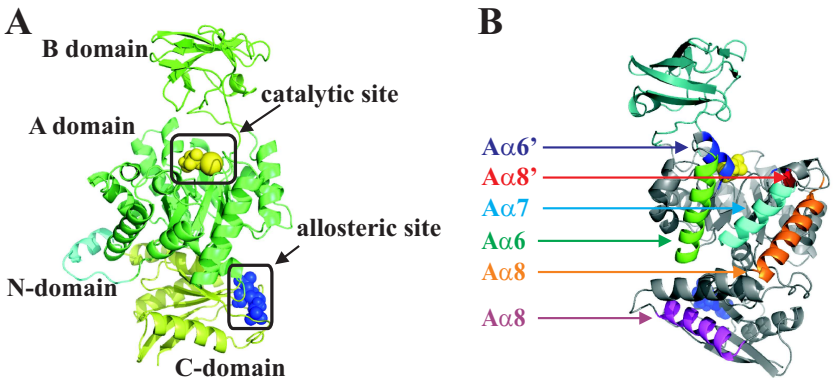

Figure 7.2: Structure of one of the four subunits of pyruvate kinase. A: illustration of the four domain (three large domains A–C and one very small N-terminal domain) architecture. Allosteric and catalytic binding sites are marked. B: the same monomer with secondary structure elements highlighted which have been found to be functionally important in kinetic experiments.

is also the activator of nearly all characterised eukaryotic PYK molecules. In mammals, there are four distinct forms, which are named the M_1, M_2, L- and R-types: M1 PYK is mostly present in the skeletal muscle, M2 PYK in many tissues such as kidney, intestine, lung fibroblasts, testis, and stomach, L-PYK mostly in the liver, R-PYK exclusively in red blood cells. These forms all consist of identical subunits (see Section 7.2.1) of about 60 kDa, but differ in enzymatic properties as well as in regulation of their gene expression [190].

7.2.1 Crystal Structures

Overall architecture

The molecular architecture of the enzyme is very complex. It comprises tetramers of four identical subunits, each monomer encompassing three domains (cf. Fig. 7.2A); the A domain shows $(\beta/\alpha)_8$ barrel topology, the small B domain, characterised by an irregular β barrel, the C domain has an α/β topology. A fourth small N-terminal domain, absent in bacterial PYK, is formed by a helix-turn-helix motif.

Domain A (residues 1–86, 186–357 and 424–435) is the largest of the three major

domains. Its $(\beta/\alpha)_8$-barrel is characterised by three additional α-helical segments located on loops 6, 7 and 8 (named Aα6', Aα7', Aα8') at the C-terminal side of the eight-stranded parallel β-sheet, play a central role in catalysis and allosteric regulation (cf. Fig. 7.2 B).

Domain B (residues 87–185), adjacent to the C-terminal side of domain A and inserted on loop 3, forms a lid covering the $(\beta/\alpha)_8$-barrel. Domain B consists of a mixed β-barrel with only one short α-helix. It interacts very loosely with domain A, connected by only one interdomain hydrogen bond (hydrogen bonds not shown in the figure). In contrast, domain A interacts tightly on its N-terminal side with domain C (residues 352–470), which displays an α/β structure with $\alpha\beta\alpha\beta\alpha\beta\alpha\beta\beta$-topology.

The active site is located between the A and the B domain in each of PYK's four subunits (yellow ligand in Fig. 7.2A). The binding site for the allosteric activator, FBP (blue ligand in Fig. 7.2A), is completely contained in the C domain .

In the functional enzyme, the four subunits are closely associated to form a tetramer of 222 symmetry. The 222 symmetry is exact only for the A and C domains, whereas the B domains of subunits 1 and 3 are not related to each other by exact twofold axes because of their different orientations. Extensive interactions mainly between the A and the C domains of two opposite subunits stabilize the assembly, whereas the B domains do not take part in any inter-subunit interaction. Intermolecular contacts along the larger interface are mostly contributed by residues belonging to helices Aα6, Aα7 and Aα8. The most relevant structural elements building this interface are strands Cβ5 of the two opposite subunits which, by running antiparallel to each other, extend the central β-sheet of the C domains to generate a 10-stranded intermolecular β-sheet.

7.2.2 PYK allosteric activation – experimental results

Allosteric regulation of PYK is a critical feature of cell-cycle control in rapidly proliferating tissues, because PYK controls both the consumption of metabolic carbon for biosynthesis and the utilization of pyruvate for energy production. Defects in regulation of PYK activity have been shown to be involved in the growth of malignant tumor cells (see [188] and references therein).

PYK is a typical allosteric enzyme of the K-type, i.e. an enzyme, whose Michaelis-Menten constant K_m is being altered by effectors. This is in contrast to the other class of allosteric proteins, V-type, which have their maximum reaction rate $V_{\max}$ altered upon activator binding. Crystallographic and mutagenesis studies have revealed the

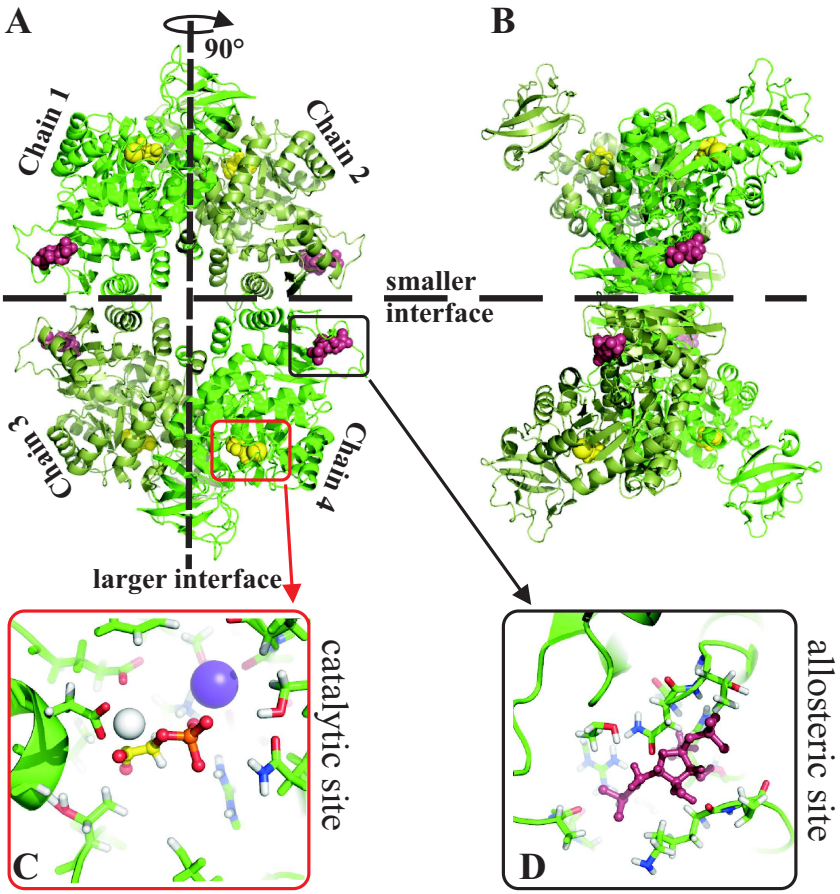

Figure 7.3: Structure of *l. mexicana* pyruvate kinase. A, illustration of the tetramer architecture. The four subunits (monomers 1-4) are separated by two interfaces (dashed lines). B, the same structure rotated by 90° illustrating the X-shape of the tetramer from the side. C, red box, the catalytic binding site with natural ligand phosphoglycolytic acid (PG, a PEP replacement) with adjacent K^+ (purplish) and Mg^{2+} (white) ions. D, black box, the allosteric site with the natural ligand fructose-2,6-bisphosphate (purple) docked inside.

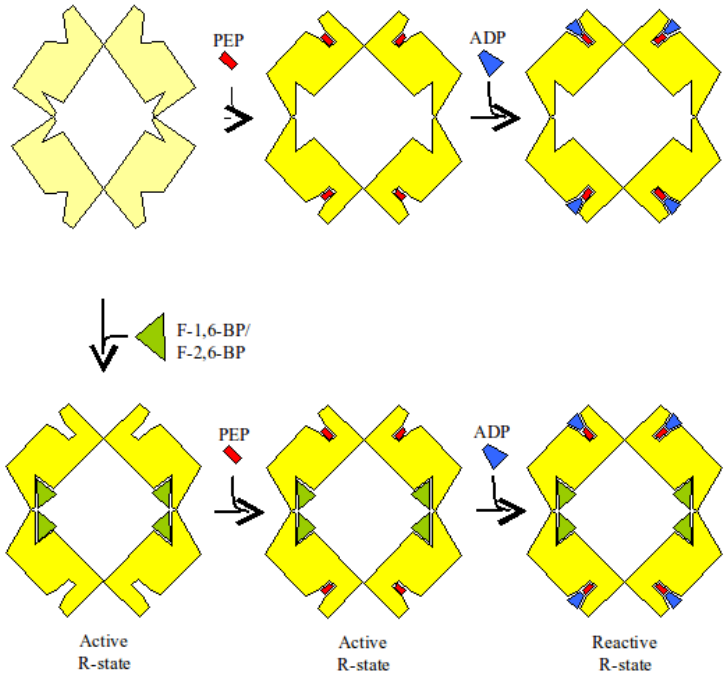

Figure 7.4: Sketch of the allosteric activation of PYK; homotropic vs. heterotropic activation [191]. Top row, homotropic activation: Inactive T-state PYK can be activated through the binding of PEP to the catalytic site. However, PEP has a low affinity for the active site of the inactive enzyme ($S_{0.5} = 1.44\,\mathrm{mM}$) [192]). The binding of PEP allows the binding of ADP at the active site. Bottom row, heterotropic activation: Inactive T-state PYK can also be activated through the binding of F-1,6-BP or F-2,6-BP to the allosteric site. The activated enzyme has a much greater affinity for PEP ($S_{0.5} = 0.2\,\mathrm{mM}$) [192]) than the inactive enzyme.

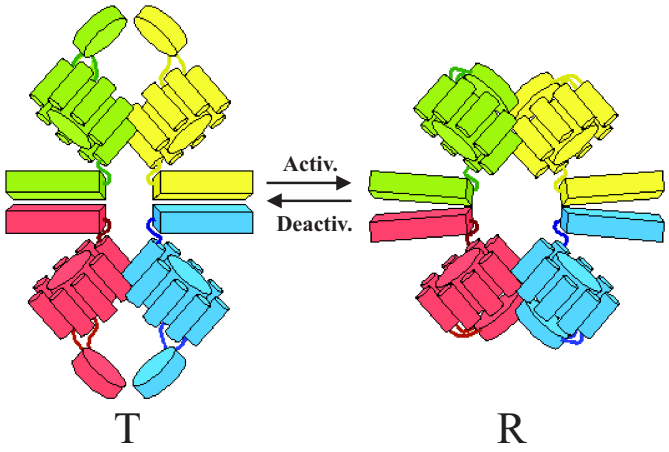

Figure 7.5: Schematic layout of pyruvate kinase illustrating structural differences between inactive T-state (left) and activated R-state (right) [191]. Different colours illustrate different subunits. The barrel architecture of the A-domains (central to each of the subunits) is sketched, B-domains are symbolized as lids, C-domains as base planks. The quarternary structure of the T-state is much narrower and higher than the one of the R-state.

general features of the allosteric transition mechanism. The enzyme displays a dramatic conformational change in going from T- to the R-state. The allosteric signal is transmitted across the long distance ($> 20\,\text{Å}$) separating the allosteric binding site from the catalytic center. The exact mechanism of the allosteric transition is not known in detail, since not PYK had been so far crystallised in both T and R-state. Even though allosteric effects at the active site have been characterized and the FBP binding site is known [188], very little is known about the propagation of the allosteric signal from the FBP binding site to the active site, because PYK had so far not been crystallized in both the T and the R states. Two mechanisms of allosteric transition have been proposed.

Network mechanism

Friesen *et al.* [193] proposed a network mechanism to explain allostery in PYK. They proposed that K421 in rabbit muscle PYK hydrogen binds with Y443 of the neighbouring subunit across the C-C interface. They then proposed that this tyrosine residue is part of a hydrogen bond network that transmits an allosteric signal from the C-C interface to the active site. Results by Fenton and Blair [194] support this theory. This network

mechanism is in contrast to the rotating rigid domains mechanism suggested by Mattevi *et al.*

Rotating rigid domains

Comparing a T-state *E. coli* PYK structure with the R-state of rabbit muscle M_1 PYK, which is not allosterically regulated and is thus thought to be an allosterically locked 'natural' mutant adopting an active R-like conformation [195], Mattevi *et al.* [189, 196] proposed a combination of two kinds of movements: (i) intra-subunit rotation of the B and C domains (by 17° and 15°, respectively) within every subunit, and (ii) inter-subunit 16° rotation. The activation process therefore seemed to involve a combination of domain and subunit rotations, whereas the structure of the individual domains is not significantly altered, implying that they rotate as rigid bodies and are connected to each other by highly flexible hinges.

However, this proposed mechanism suffered from the fact that the PYK structures compared were not only from different species, but one was allosteric and the other was not. All known PYK isoenzymes share high similarity in their primary sequences. However, extrapolating effects of mutations in one isoenzyme to other isoenzymes may lead to severe errors, since minor sequence differences may cause large changes in stabilization, enzymatic, and thermodynamic properties [194]. The model of allosteric activation was improved through a comparison of the structures of inactive T-state *L. mexicana* PYK with active R-state *S. cerevisiae* PYK [197], allowing to better separate the structural differences that are consequences of the inherent divergence between eukaryotic and prokaryotic proteins from the conformational changes associated with the allosteric transition.

The mechanism of PYK regulation has also been the subject of many mutagenesis experiments [194, 198]. The general picture emerging from the mutant analysis is that the intersubunit interactions at the A/A' and the C/C' interfaces and the interdomain interactions at the A/B interface are key to determining the allosteric response and to defining the distribution of the conformations between active and inactive states. Moreover, the mutagenesis analyses combined with the crystallographic data provide evidence for the idea that the T and R forms correspond to ensembles of conformations characterized by the rotational flexibility of the B domain. The key functional role of the A/A' interface has been highlighted by a mutation, which targets a residue located in the core of the A/A' interface, producing an enzyme the retains a stable tetrameric state

but almost entirely lacks enzymatic activity [199]. Conversely, none of the mutations targeting residues in the A/C interface alters the enzyme's allosteric properties. Thus, this domain interface appears not to be involved in the transduction of the allosteric signal; rather it seems to be important for the stability of the domain assembly within the enzymatic subunit [194].

All of the above proposals were based on comparisons of crystal structures and enzymatics of mutants. Information on dynamics as well as considerations of intermediate states, in contrast, were lacking. In this study, extensive molecular dynamics simulations were used to elaborate a model of allosteric *L. mexicana* PYK regulation. The simulations were started from and compared to three crystal structures of two different PYK isoenzymes. First, the T-state of *L. mexicana* PYK, second, the fully activated *yeast* PYK R-state, and, third, the recently solved crystal structure of a putatively not fully activated *L. mexicana* R-state. Here, the fact that the crystallisation precipitant's sulphate moieties occupied the binding sites of the phosphate groups of both fructose-2,6-bisphosphate and PEP, displacing all native ligands, triggered speculations that this structure might not be fully active. This notion was corroborated by observations that both the orientation of the monomers with respect to each other (quarternary structure) and the intra-subunit structure (tertiary structure) were different from R-states of different species, in particular of the R-state of *yeast* pyruvate kinase.

We first aimed at obtaining an ensemble of inactive T-state structures and active R-state structures, respectively, followed by examinations of the influence of ligand (un-)binding and coupling between the subunits. Finally, we characterised the homotropic activation pathway of PYK by comparing changes of dynamics of differently ligated intermediate states.

7.3 MD simulation details

T-state structure

The T-state structure of wild type *L. mexicana* pyruvate kinase at 2.35 Å was taken from the protein database (entry 1PKL). Two tetramer structures were available (first tetramer chains A–D, second tetramer chains E–H) and only one of the eight chains were complete without any regions of poor electron density (chain H). Non-resolved regions of other chains were modelled according to the corresponding regions of Chain H using the

WHATIF suite [200] and the editconf and g_rmsd modules of the GROMACS simulation suite [113, 114].

R-state structures

The (3.3 Å) crystal structure of the 'quasi-active' R-state of *L. mexicana* PYK was recently crystallized (pdb code 1E0V) and described [201]. None of the present subunits were fully resolved, such that non-resolved loop regions were modelled according to the resolved loop regions of Chain H of the active T-state structure (see above). The R-state structure of *yeast* PYK was taken from the protein data base (entry 1A3W). Since this structure served only as a reference point and no MD simulations were carried out for this isoform, no further refinement was carried out here.

General simulation parameters

All simulations were carried out with the Gromacs simulation suite [113, 114], using the OPLS all-atom force field [64] and periodic boundary conditions. NpT ensembles were simulated, with the protein and solvent coupled separately to a 300 K heat bath ($\tau_T = 0.1$ ps) [69]. The systems were isotropically coupled to a pressure bath at 1 bar ($\tau_p = 0.1$ ps) [69]. Application of the Lincs [31] and Settle [30] algorithms allowed for an integration time step of 2 fs. Short-range electrostatic and Lennard-Jones interactions were calculated within a cut-off of 1.0 nm, and the neighbour list was updated every 10 steps. The particle mesh Ewald (PME) method was used for the long-range electrostatic interactions [72], with a grid spacing of about 0.12 nm.

Ligand force field

The natural ligand phosphoglycolyic acid (PG), which is used in all structures of all PYK isoforms crystallised so far as phosphoenol pyruvate replacement, and fructose-2,6-bisphosphate (FBP), were not available in the OPLS force field. The atomic partial charges of both FPB and PG were derived from quantum chemical calculations at the B3LYP/6-31+G* level using the CHELPG electrostatic potential fitting scheme [115]. The quantum chemical calculations were carried out with Gaussian03 [116]. All non-bonded parameters are given in Appendix B. Bonded parameters were taken from Reference [117].

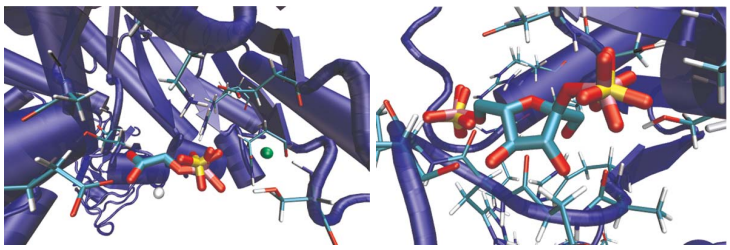

Figure 7.6: Docking of natural ligands PG (left) and FBP (right) into the binding sites of the putative R-state. The phosphate groups of FBP and PG overlap with the crystallisation sulphates proposed to have replaced PG and during the crystallisation process. For PG with only one phosphate group, binding site information of *yeast* PYK was additionally used to allow for unambiguous binding.

Docking of PG and FBP

PG and FBP were docked into the corresponding binding sites such that the phosphate moieties overlapped with the sulphate ions present in the crystal structure (cf. Fig. 7.6). Additionally, structural information from the *yeast* PYK binding sites, where all natural ligands were resolved, were used to guarantee correct distances to binding site residues. A short MD (500 ps) simulation was used to equilibrate the ligands within the binding sites.

Energy Minimization and Equilibration of the Solvent

Prior to the free MD simulations, the systems were solvated with explicit TIP4P water within a dodecahedral box of 16.9 nm length such that the minimum distance of the protein from the box boundaries was 1.2 nm. Sodium and chloride ions were added ($c = 0.15\,\mathrm{mol/l}$), and the systems were energy minimized for 1000 steps using steepest descent. The solvent was then equilibrated for 500 ps with positional restraints on the protein heavy atoms (force constant $1000\,\mathrm{kJ\ mol^{-1}\,nm^{-2}}$). The total system comprised roughly 500.000 atoms. Total simulation time of all studied systems was about 500 ns.

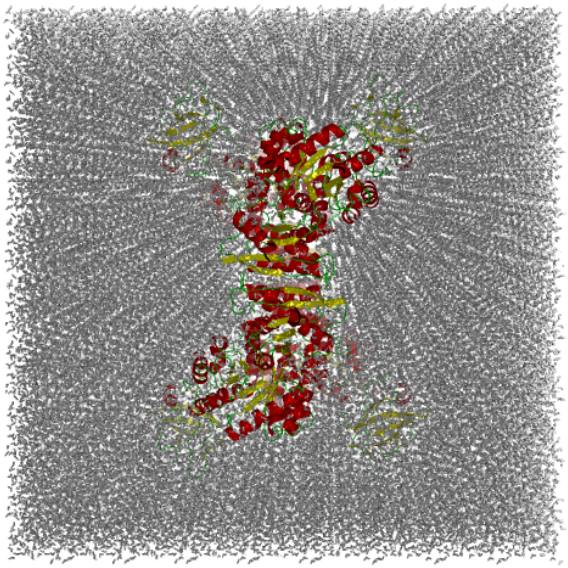

Figure 7.7: Pyruvate kinase simulation box (cubic only for illustration purposes) totalling 500.000 atoms. Protein coloured, explicit solvent grey.

7.4 Results

7.4.1 Identification of T- and R-state

For the examination of allosteric activation pathways it is important to identify the active and inactive state of the same isoenzyme. In the first part of this study, we addressed the question whether the new crystal structure of the quasi-R state (cf. Sec. 7.3) has to be considered a fully or a partially activated structure. This question was raised [201], because the crystal structure showed non-native ligands inside all binding sites, and both quarternary and tertiary structure appear not close enough to previously crystallised fully activated R-states of other isoenzymes. Accordingly, it was proposed that this structure does not belong to the fully activated but to a trapped intermediate state, whereas the true *L. mexicana* R-state was proposed to show larger structural similarity to the R-state of yeast. Two tests were carried out to test this proposal. First, the influence of native vs. non-native ligands on the dynamics of the quasi-R state structure was compared and, second, the system was driven towards the fully activated *yeast* R-state structure, followed by an analysis of the sampling behaviour after removal of constraints.

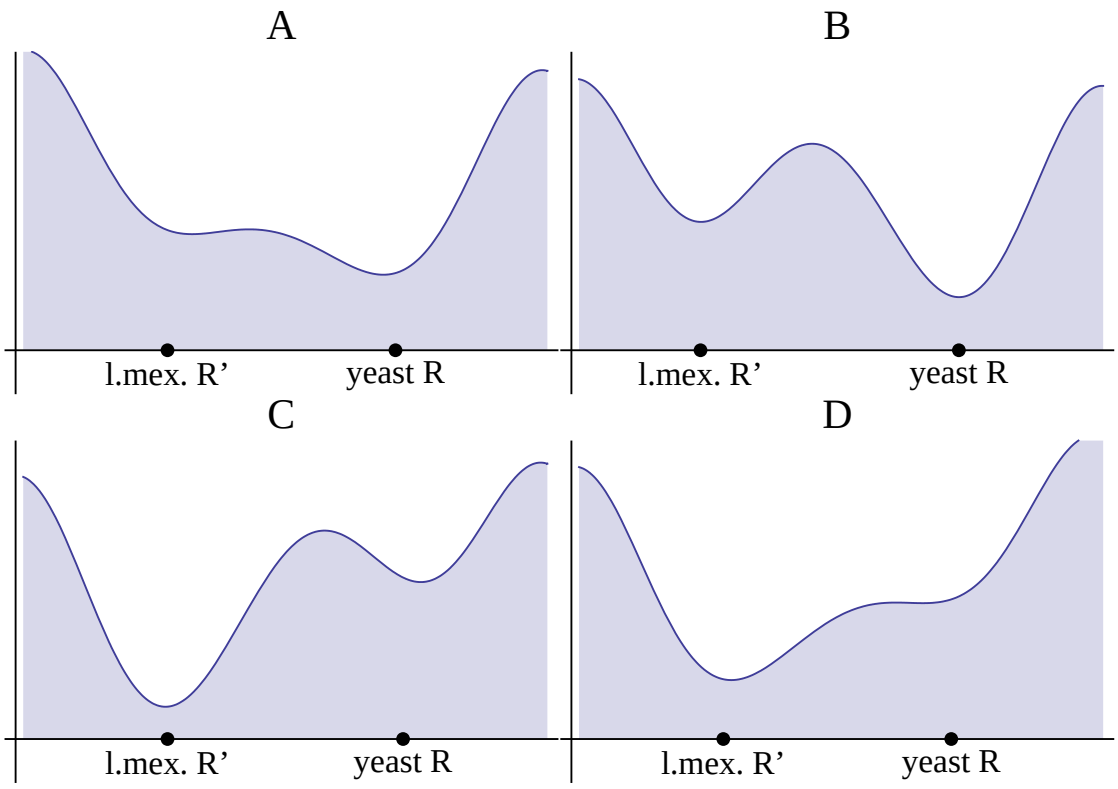

Figure 7.8: Sketched free energy surfaces. *l.mex.* R': recently crystallised (partially activated?) *L. mexicana* R-state structure, *yeast* R: fully activated *yeast* R-state. A: The *L. mexicana.* R'-state is energetically unfavourable. The fully activated state close to the *yeast* R-state (right), however, is readily accessible; B: Like A, except *L. mexicana.* R' is trapped inside an energy well separated from the more favourable fully activated conformation by a high energy barrier; C: The *L. mexicana.* R'-state is energetically more favourable than the region close to to the *yeast* R-state. Stimulated transition over the energy barrier would result in a new trapped ensemble; D: like C, except a stimulated sampling of the region close to the *yeast* R-state would result in an immediate back-transition.

How well does sulphate really mimic native ligands?

To test how well the sulphate ions present in the R-state mimic the native ligands PG and FBP, we compared the sampling of the 'quasi'-R state with SO_4^{2-} to the same structure with native ligands, which were docked into the binding pockets as described in methods. If the sulphate ions are indeed the reason for PYK to have failed to fully activate, pronounced differences of dynamics or even a straight structural rearrangement to a fully activated conformation are expected to be seen. In Fig. 7.8A, where a free energy landscape is sketched, this would correspond to a transition over a very low energy barrier from the *L. mexicana* quasi R-state (*l.mex.* R', left) to the fully activated R-state close to the *yeast* R-state (*yeast* R, right).

Figure 7.9 shows the sampling behaviour projected in a quarternary-tertiary (Q-T) PCA coordinate system, which shows the quarternary change on the x-axis and the tertiary change on the y-axis. The quarternary axis represents the PCA eigenvector (based on the 1892 C-α atoms that were resolved in all crystal structures), which describes a transition from the *L. mexicana* T-state to the *yeast* R-state with constrained *intra*-subunit motions. Conversely, the tertiary axis is the PCA eigenvector which describes the same transition with constrained *inter*-subunit motions.

The projections of the three crystal structures on these two eigenvectors are symbolised as red squares. The *yeast* PYK R-state (1A3W) exhibits the strongest domain closure (tertiary coordinate) and the smallest angles between subunits (quarternary coordinate). It can thus be viewed as the most R-like reference point. The lower left red square represents the T-state structure of *L. mexicana* PYK (1PKL), exhibiting the most T-like behaviour. The new crystal structure of the *L. mexicana* 'R'-state (1E0V) is situated in the middle. For both *L. mexicana* isoforms, MD simulations were carried out; the T-state without any ligands, the *L. mexicana.* 'R'-state with SO_4 and PG+FBP, respectively. The *yeast* R-state served as a reference point only; no simulation was started from there.

As can be seen, the ensembles pertaining to T- or R-state do not overlap. Instead, a clear gap is observed between the two ensembles. Furthermore, for the R'-state, both trajectories, with sulphate and native ligands, respectively, sample the same region of configurational space in the first place, travelling into the R-direction of the tertiary coordinate but slightly in T-direction on the quarternary coordinate. The small gap between crystal structures and ensemble is due to non-recorded sample points during the equilibration period. Interestingly, the docked native ligands triggered slightly stronger

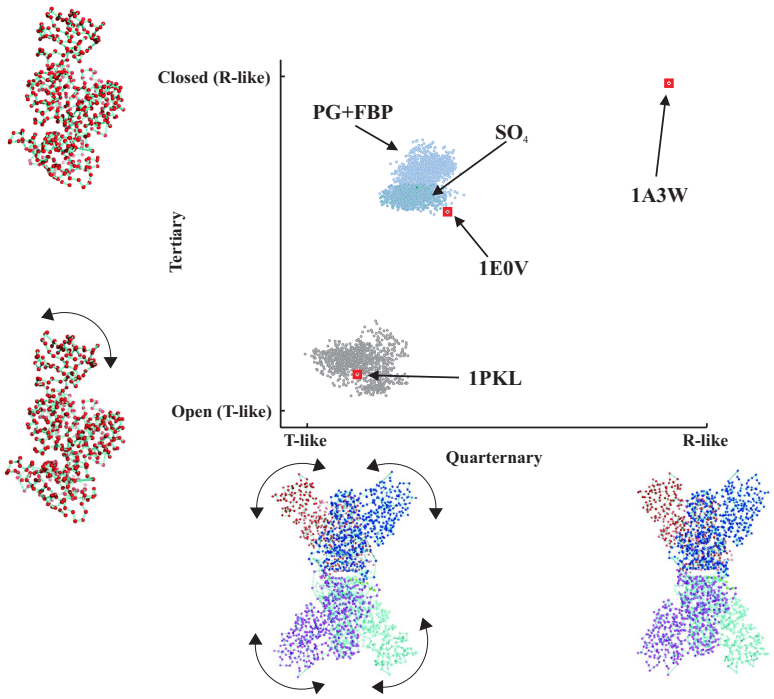

Figure 7.9: Coordinate system with a quarternary coordinate (x-axis) and a tertiary coordinate (y-axis). Red dots correspond to crystal structures, sample point clouds to equilibration trajectories. Lower left: T-state PYK, middle: SO_4-labelled – sulphate ions as obtained from the crystal structure left inside the bindint sites, 'PG+FBP'-labelled: native ligands docked inside all binding sites. Cartoon representations outside the graph illustrate the main movements represented by tertiary and quarternary coordinate.

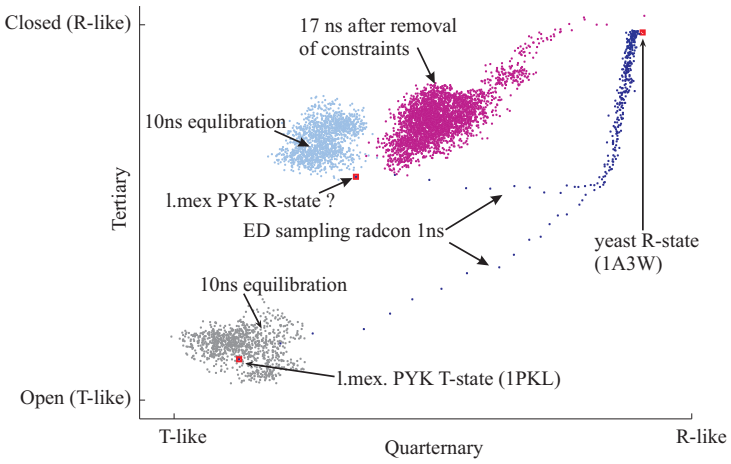

Figure 7.10: Essential dynamics stimulated transition (blue) from both *l. mexicana* states (left) towards the putative fully activated *yeast* R-state (upper right) and backtransition after removal of constraints (pink).

R-like sampling on the tertiary coordinate indicating marked domain closure, albeit not to the extent seen in *yeast* PYK. Overall, no pronounced movement into the 'full' R-direction (upper right) was seen. Thus, the situation skeched in Fig. 7.8A is not present.

This allows two conclusions. Either the crystallised pseudo-R state is not an intermediate but a fully activated state or the timescales of the simulations did not suffice for the transition to the fully activated R-state. The latter case would occur if the energetically more favourable fully activated state would be separated from the ensemble around the crystallised intermediate state, as depicted in Fig. 7.8B.

Is yeast R-state close to the "real" L. mexicana. R-state?

The possibility that an energy barrier prevents the system from reaching a potentially energetically more favourable fully active state during the relatively short simulation time has to be addressed.

To stimulate a transition over a possible energy barrier, essential dynamics simulations (ED, cf. methods Chapter 2) were carried out driving the system towards the *yeast* R-state, the most R-like point in the Q-T coordinate system. Radius contraction, as described in methods, was carried out on the first 1000 of the total 5676 eigenvectors obtained after a PCA of the the ensembles presented in Fig. 7.9. Only the 1892 C-α

atoms that are resolved in all three crystal structures were included into the analysis. Even though the radius contraction algorithm does not strictly force the system into the desired direction (cf. methods), and is thus a relatively soft sampling enhancement, the transition (blue) of both T-state (with native ligands docked into the previously empty binding sites) and *L. mexicana.* 'R'-state (with native ligands docked into the previously sulphate occupied binding sites) towards the *yeast* R-state was complete after less than 1 ns, as can be seen from Fig. 7.10, resulting in an overall RMSD of around 0.1 nm (data not shown). After the transition was complete, the simulation was continued under ED-constraints for another 3 ns. In the next step, the constraints were removed and a free 'relaxation' MD was started (Fig. 7.10). Since the relaxation pathway of the R-state-started ED trajectory overlapped with the one started from the T-state, the latter was not plotted separately for clarity.

The fact that the ED activation took only 1 ns despite only a relatively soft sampling enhancement shows that no high energy barriers had to be overcome. This observation would, in principle, also agree with Fig. 7.8A. However, removal of constraints after complete transition resulted in immediate return (Fig. 7.10) to the *L. mexicana.* 'R'-state ensemble within 17 ns. This observation rules out the possibility that favourable energy wells exist for *L. mexicana* PYK structurally close to the *yeast* R-state (as depicted in Fig. 7.8B and C). In contrast, all observations are in agreement with Fig. 7.8D, where the *L. mexicana* 'quasi' R-state sits in a favourable energy well and the region structurally close to the *yeast* R-state is rather unfavourable. Furthermore, from the fact that this backtransition also happened when the transition started from the T-state (with native ligands bound), we conclude that the energetically most favourable state for *L. mexicana* PYK with ligands bound is indeed close to the quasi-R state structure.

7.4.2 Do the ligands induce immediate structural rearrangements?

Our results so far do not completely rule out the possibility that the true *L. mexicana* R-state is similar to neither the *yeast* R-state nor the available *L. mexicana* quasi-R state, but rather to a structure not considered in the simulations at all. To further explore the characteristics of the free energy landscape which underlies the dynamics of PYK, in the following sections we examine sampling pathways as response to ligand binding and unbinding.

In this section, the question is addressed whether or not the presence of ligands

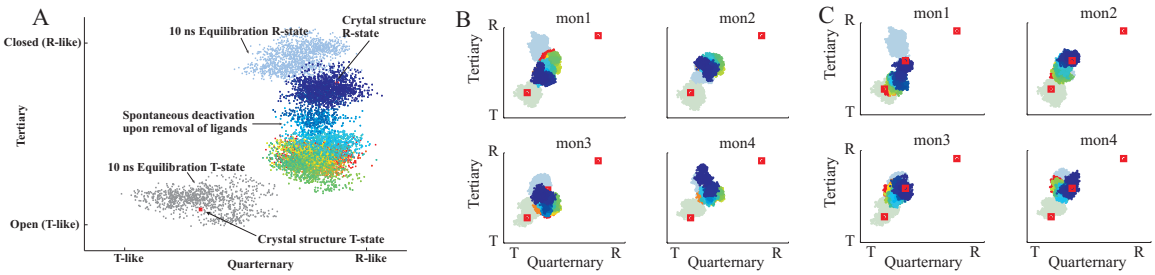

Figure 7.11: A: Spontaneous R → T deactivation upon removal of ligands. The deactivation starts immediately after removal of the ligands and is almost complete after 35 ns. As a reference, the equilibration ensembles of T- and R-state, as well as coordinates of the crystal structures, are also given. B and C: analogous deactivation processes (rainbow scatter) like (A) started from different snapshots of the R-state equilibration trajectory (bluish). In both B and C, the dynamics of the four subunits are plotted separately showing pronounced different sampling behaviour.

inside the PYK binding sites has readily observable structural consequences on the timescales observable by MD simulations. From the PYK millisecond turnover number [192], one activation/deactivation process would rather be expected to require hundreds of microseconds. One can, however, hope to see the beginnings of such a transition or pronounced changes of dynamics. Based on the assumption that activation and deactivation pathways are similar, the deactivation pathway was chosen first, due to the much simpler technical implementation – the ligands were removed and the reaction of the system observed.

Fig. 7.11 shows the observed pathways. In Fig. 7.11A, two 10 ns equilibration trajectories are shown, grey for the T-state trajectory with no ligands inside the binding sites, blue for the *L. mexicana* R-state trajectory with ligands bound. The respective crystal structures are depicted again as red squares. Note that the *yeast* R-state is not considerated here. After removal of all ligands from the binding sites, a transition towards the T-state ensemble could be seen (rainbow colours, where the colour indicates time information from blue to red), which started immediately after removal of the ligands and was almost complete after 35 ns.

To rule out the possibility that our observation was merely by chance, we carried out three more analogous deactivation tests starting from different snapshots from a 30 ns equilibration trajectory of the R-state ensemble (with native ligands bound). Additionally,

as a second step, native ligands were docked also into the T-state binding pockets and the resulting pathways were observed.

Two of those further tests are shown in Fig. 7.11B and C, respectively. Here, in contrast to A and all aforementioned plots, the observed projections of the trajectories on the tertiary coordinate are shown separately for each of the four subunits (referred to as mon1–mon4) and not as an averaged tertiary coordinate. Also in contrast to A, the *yeast* R-state is given as a red square (upper right corners in all subplots).

Interestingly, we found that the subunits show pronounced dynamics differences already in the R-state equilibration trajectory (light blue). In particular, in contrast to the other subunits, subunit 1 showed a clear domain closure. After removal of the ligands out of the R-state (rainbow colours, where the colour again indicates time information from blue to red) pronounced spontaneous movements towards the T-state ensemble could be observed. Here, too, the movements of the four subunits were markedly different. For example, whereas subunit 1 (mon1) in C performed a quick and complete transition towards the T-state ensemble, subunit 2 in B did not leave its native R-state ensemble (light blue) at all. Similarly, subunit 4 (mon4) in B initially performed a domain closure motion, upwards in the figure, before reversing its direction towards the T-state. But in C, the same subunit initially started towards the T-state but reversed the direction after half of the transition.

A similar situation was seen for the reverse case, the T-state structure with native ligands bound. For this case, a 20 ns trajectory was simulated (light green in Fig. 7.11B and C). While, like above, the dynamics of the subunits appeared to be different, no clear overall transition was seen. Only subunit 3 appeared to start a T $\rightarrow$ R transition, wheras the other three did not leave their native T-state ensemble.

To sum up, it was seen that absence of ligands introduced markedly different PYK dynamics compared to the same system with ligands bound. Immediately after simulation start, in all observed cases the systems showed a clear tendency to leave the native ensemble and to approach the T-state both in terms of the quarternary coordinate and in terms of the overall tertiary coordinate. Conversely, binding ligands into the empty binding sites of the T-state structure did not (on the timescales accessed by our MD simulations) result in T $\rightarrow$ R transitions.

While, thus, the overall reaction of the system upon ligand removal indicates the existence of an induced fit mechanism, the situation is somewhat more diverse if subunit-wise tertiary coordinates are employed rather than averages over the tertiary coordinates

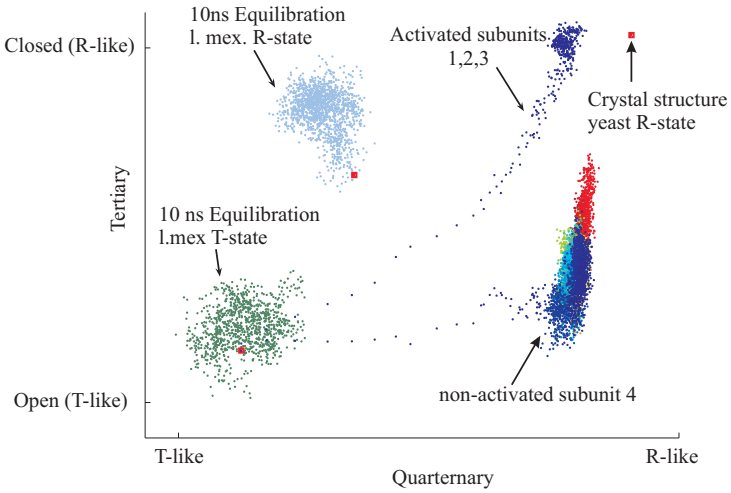

Figure 7.12: Coupling of the tertiary structures of different subunits. ED stimulated activation of subunits 1–3 from the T-state ensemble (green) induced the fourth monomer to follow suit within 30 ns.

of all subunits. A pronounced diversity then becomes evident, which brings up the question, whether or not subunit dynamics are influenced by neighbouring subunits' dynamics at all.

7.4.3 Are the tertiary structures of the subunits coupled?

Whereas the previous section established that the ligands do have considerable influence on both the tertiary and the quarternary architecture of the PYK tetramer, it could not be shown that the internal structure of each subunit depends on neighbouring subunits' conformation. From the results so far, it would be entirely possible that the conformation of each subunit merely reacts on ligand binding or unbinding without being sensitive to the neighbours' binding situation and conformation at all. This, however, would then not explain the cooperative effect.

To examine the structural influence of inter-subunit configurational coupling, we again carried out radius contraction essential dynamics simulations. In contrast to the ED simulations of Sec. 7.3, only the 1419 C-α atoms that were resolved in both *L. mexicana* PYK and *yeast* PYK and belonged to subunits 1, 2 or 3, respectively, were subjected to ED enhanced sampling. Within the binding sites of these three subunits, the native ligands FBP and PG were bound. They were then driven towards the *yeast* R-state. The

fourth subunit's binding sites, in contrast, were left empty and this subunit was also not subjected to ED, but left free to move. If the subunits' conformations influenced their neighbours, the unperturbed fourth subunit should show markedly different dynamics after its neighbouring subunits were subjected to ED as compared to a completely unperturbed tetramer.

Fig. 7.12 shows that the ED activated subunits 1, 2 and 3 (dark blue), which were started from the T-state (red square lower left) equilibration ensemble (light green), reached a tertiary coordinate similar to the *yeast* R-state (red square upper right) within less than 5 ns. The quarternary coordinate, which is a property of the whole tetramer, also showed such a fast activation procedure. Even after this completed activation, the ED constraints were kept switched on throughout the whole subsequent simulation, when the dynamics of the fourth subunit were observed. This unperturbed fourth subunit (rainbow colours from blue to red represent time information) exhibited a similar domain closure movement like the activated three other subunits, i.e. an upwards tendency on the tertiary coordinate. However, the closure motion of the fourth subunit did not reach the extent of the perturbed subunits, after reaching a similar tertiary coordinate like the *L. mexicana* R-state structure (red square middle left). Interestingly, the initial closing motion stopped after about 15 ns (change from blue to green). During the following 12 ns, a pronounced opening motion took place reaching almost the fully opened starting point before the closing carried on to reach its final position around 30 ns (red).
Such structural rearrangements could neither be observed for empty subunits nor for holo subunits (with bound ligands) in the previous free MD simulations (cf. Sec. 7.4.2). The activation of the three subunits appeared to be connected to the fast structural rearrangements seen for the non-perturbed subunit. We conclude that the activated conformations/dynamics of the activated subunits 1, 2 and 3 significantly flattened the free energy landscape underlying the dynamics of subunit 4, such that the T $\rightarrow$ R-transition of subunit 4 was facilitated.

However, whereas this experiment showed that intra-subunit configurations are influenced by neighbouring subunits, the fact that the conformations of three of the four subunits were constrained throughout the whole simulation has to be considered a severe interference with the natural dynamics of the protein. All simulations with unperturbed dynamics so far compared the dynamics of the fully ligated with the completely unligated PYK. Accordingly, it is as yet unclear, how the activation from the inactivated empty T-state PYK to the activated R-state PYK actually takes place. As a start, the following

section aims at following the pathway of homotropic PYK activation, which is sketched in Fig. 7.4.

7.4.4 Homotropic activation of pyruvate kinase

We aim at characterising the homotropic activation of PYK in terms of changes of dynamics of differently ligated PYK configurations. The advantage of examining homotropic activation rather than heterotropic activation is the fact that here the number of intermediate steps is much lower; and it is the intermediate states which have been proposed to carry the clue to understanding allosteric activation processes.

In kinetic experiments, homotropic activation was observed to take place very slowly, due to a very low affinity of the PEP/PG ligands to the catalytic site of the inactive T-state (cf. Fig. 7.4). This first obstacle was overcome by docking PG with accompanying ions properly into the catalytic site, as described above. Due to the symmetry of the tetramer, a set of six 50 ns simulations of differently ligated T-state PYK molecules was sufficient to examine all different ligation states; one single-liganded, three double-liganded (one cross-over, one small-interface neighboured, one larger interface neighboured), one triple-liganded and one full-liganded. Over the whole 300 ns trajectory, a full-residue (including side-chains, but excluding hydrogens, which are not expected to significantly change the result; altogether 45,000 coordinates were examined) PCA was conducted yielding a set of eigenvectors which best represent the common collective dynamics to the system. Each of the first few principal components thus obtained turned out to represent well the collective motions of exactly one of the ensembles underlying the PCA.

The residue-wise RMSF plot (Fig. 7.13 1a) shows the distinct distribution of highly flexible regions along the eigenmode pertaining to the single-ligated PYK ensemble. Above-average flexible regions are marked red. This docking of only one ligand into the tetramer yielded a fast T $\rightarrow$ R transition not only in this particular ligated subunit (holo, Fig. 7.13 1b left) but also of the empty small-interface neighbour (holo, Fig. 7.13 1b right), which was even more pronounced than for the ligated subunit. This surprising result can be understood from the distribution of the active regions along this collective coordinate (obtained from Fig. 7.13 1a), which are mapped into a cartoon representation of PYK. The ligated subunit (green) exhibits high activity around its own allosteric site (allosteric effector FBP in grey for better orientation only, it was absent in the simulations) as well as around the smaller interface (Fig. 7.13 1e), where crucial regions for signal transductions

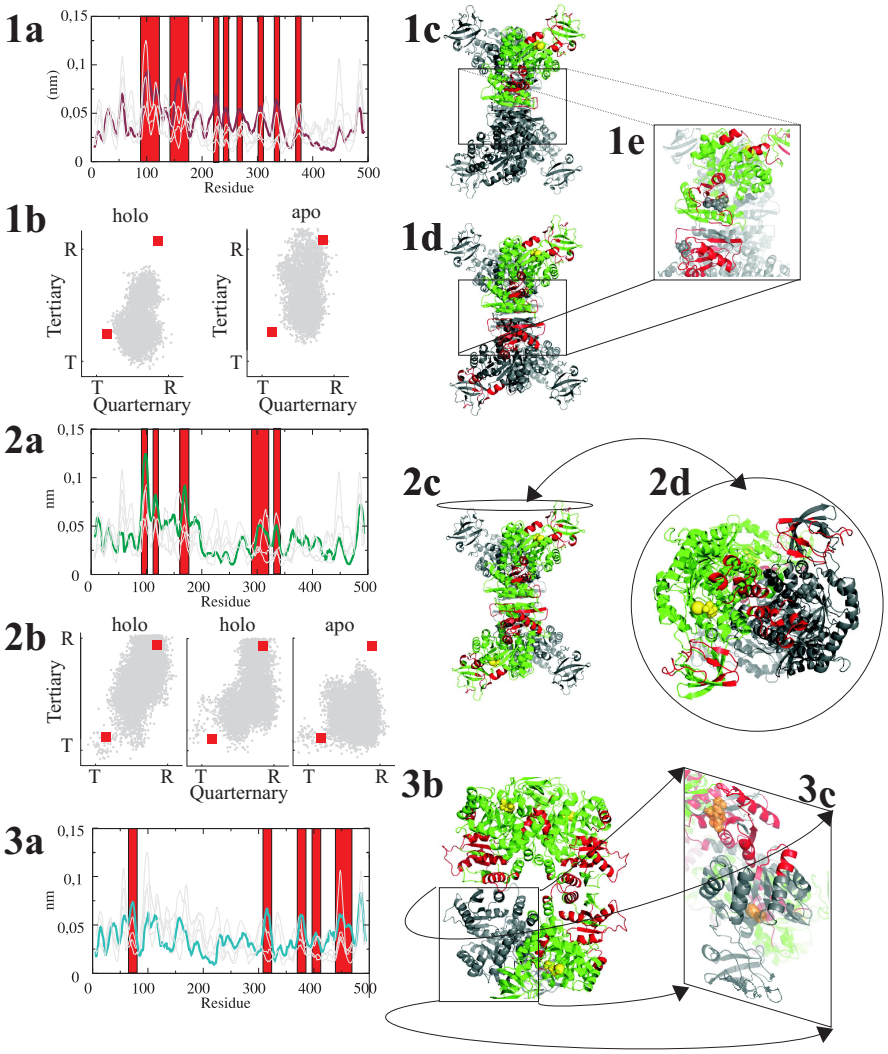

Figure 7.13: Collective motions observed during the homotropic activation of PYK. 1a: Distribution of especially flexible regions (red) in PYK upon docking of only one ligand in the tetramer as obtained by RMSF. 1b: The ligation of one subunit yields a T → R not only in this particular subunit (holo, left) but also of the non-ligated small-subunit neighbour (apo, right); 1c: the flexible regions obtained from and marked red in 1a are highlighted red in a cartoon representation of PYK in the liganded subunit (side view). The ligated subunit is green, empty subunits grey. PG in yellow. The collective vibrations highly affect the allosteric binding site of the subunit, and the region around the small subunit. The larger interface is unaffected. 1d: The same flexible regions marked in the smaller-interface neighbour. Regions needed for T → R transition are highly affected; 1e: zoom into the smaller-interface region illustrating the signal jump over the interface; 2a: like 1a, except for ligands in smaller-interface neighboured subunits; 2b: this ligation combination introduces T → R transitions for the two ligated subunits (holo, left and middle) and also for the non-ligated larger-interface neighbour (apo, right); 2c) no collective movement is seen any more via the smaller interface; 2d) rather, the collective mode now forms a path via the larger interface to the larger-interfact neighbour's catalytic binding site, explaining 2b; 3a) like before, but for triple ligated PYK; 3b) the active regions now form a network which mostly affect the allosteric binding sites, but not the catalytic sites; 3c) focus on the empty subunit. No activity could be observed around the catalytic site.

are active. Mapping the flexible regions also in the neighbouring subunit illustrates which the T → R transition in this empty monomer likely occured. Following this activation route, the collective dynamics were examined of the double-ligated tetramer, where both of the just examined smaller-interface neighboured subunits are ligated. Here, somewhat different dynamics were observed and the active regions shifted to different areas of the protein (Fig. 7.13 2a). Also here a T → R transition could be observed not only for the ligated subunits (holo; Fig. 7.13 2b left and middle), but also for the larger-interface neighboured subunit (apo; Fig. 7.13 2b right). Mapping the active regions into a cartoon representation of the protein, like above, shows the possible reasons. While the activity around the smaller subunit showed a remarkable drop (Fig. 7.13 2c, ligated subunits again in green, empty in grey), a pronounced active pathway is seen which connects the catalytic sites of the larger-interface neighboured subunits (top view Fig. 7.13 2d). The PG ligand inside the non-ligated subunit (right, grey) is again only for orientation purposes, it was absent in the simulation. Ligating three of the four subunits again yielded a completely different activation pattern (Fig. 7.13 3a). Here, a network of active regions is spanned throughout the whole protein (Fig. 7.13 3b) concentrating around the allosteric binding sites. A quarternary T → R transition was seen, but no tertiary T → R transition of the empty subunit could be observed (data not shown), which is conceivable from the maps of active regions inside this empty subunit (Fig. 7.13 3c). The ligands FBP (top) and PG (below), absent in the simulations, were coloured orange for better visibility to help identifying the binding sites. Regions crucial for T → R transitions were unaffected by the network of activity observed for this ensemble, suggesting that no stimulation for domain motions were available. On the other hand, pronounced activity was observed around the allosteric site, suggesting that homotropic activation could result in increased subsequent heterotropic activation. Kinetic measurements to test this idea could be fairly easily carried out.

7.5 Conclusions and outlook

Our results show that, surprisingly, pyruvate kinase is regulated on time-scales which are very well accessible to atomistic simulation techniques, despite a turnover number which indicates catalytic cycle timescales on the order of milliseconds. In a coordinate system spanned by three crystal structures of two different isoenzymes, PYK dynamics was depicted as a linear combination of changes of quarternary and tertiary coordinates.

We showed that, based on PYK dynamics, the recently crystallised putative R-state of *L. mexicana.* PYK has likely assumed a fully activated R-state structure; in contrast to what was proposed based on structural comparison and the fact that sulphate ions replaced the natural ligands inside both allosteric and catalytic site. Activation and deactivation appear to follow an induced fit mechanism, exhibiting spontaneous T $\rightarrow$ R (R $\rightarrow$ T) transition upon introduction (removal) of ligands, where the activation appears to take slightly longer than the deactivation (50–70 ns vs 30 ns). The dynamics of the four subunits appears to be largely independent in the fully ligated and in the fully unligated states, in disagreement with the symmetry assumption of the MWC model, but in line with the Pauling-KNF model. Pronounced coupling not only of the quarternary structure (which is a property of the whole enzyme) but also of the internal structure of the subunits could be observed, on the other hand, for differently liganded states.

Using the homotropic activation process of pyruvate kinase as a fairly straightforward mechanism in terms of number of intermediates, we showed that ligand binding induced procounced changes of collective motions. The detailed examination of these collective motions suggest that changes in collective motions are what drives the allosteric activation of PYK. Whereas single-ligation provides means of signal transduction almost exclusively via the small-interface, simultaneous ligation of both thereby coupled subunits appears to favour collective modes facilitating signalling via the larger interface. In both cases, the quarternary structure changed from T $\rightarrow$ R, and also empty neighbouring subunits exhibited an extremely fast tertiary T $\rightarrow$ R transition, in the single-ligated state even faster than the neighbouring ligated subunit.

The surprising result that the first and second of the intermediate ligation states appear to be more active transition-wise than the triple-ligated intermediate state could provide a whole new interpretation to the fact that homotropic activation is slower than heterotropic, i.e. homotropic activation could simply lead to a dead-end during the activation process rather than be slower due to a hindered first step. This interpretation does not collide with kinetic data from titration experiments [188] which do not discriminate between dead-end and hindered-first step activation, and the large variance 2.81 ± 0.52 of the cooperativity constant n obtained from kinetic Hill-plots [192] does not oppose this interpretation either. The fact that the dead-end observed in the simulations showed remarkable activity around the allosteric sites of PYK furthermore suggests to examine the influence of FBP addition on the kinetics after homotropic activation has been completed.

An essential brick that is lacking so far in this study of *L. mexicana* pyruvate kinase

regulation processes is the proper description of the role of entropies. The large size of this enzyme requires substantial computational effort to answer the question within the minimally coupled subspace approach developed in Chapters 4–6. On the other hand, the treatment with the quasi-harmonic approximation may lead to wrong conclusions, as demonstrated before. Accordingly, this aspect of allosteric control remains as future work.

Chapter 8
8 Summary and conclusions

It is becoming increasingly clear that the old view of proteins exisiting in equilibrium between discrete conformational states is not always helpful. Experimental results from a wide range of disciplines, in particular from NMR experiments, indicate that proteins need to be viewed as complex statistical ensembles. From a thermodynamic point of view, the major question of protein science is whether and how a protein's functional properties are related to the distribution of states within this ensemble, and how a change in environmental conditions, e.g., protein or ligand binding/unbinding, or external force, affect this distribution and alter function. Within this framework, the present thesis aims at characterizing by means of atomistic molecular dynamics simulation (MD) the dynamic and functional response of two different classes of proteins to changed external parameters; titin kinase's response to external force, and the allosteric response to ligand binding of the two allosterically regulated proteins calmodulin and pyruvate kinase.

The investigations of this thesis yielded insights into protein regulation on the atomic scale. Complemented with experimental results, such as force extension curves, calorimetric measurements and crystallographic studies, our simulations significantly contributed to the understanding of protein function in terms of underlying protein dynamics. Furthermore, to quantify (changes of) protein dynamics, and of complex macromolecules in general, we developed and applied a new non-parametric method for the estimation of configurational entropies of proteins.

Mechanoenzymatics of titin kinase

The catalytic domain of the muscle protein titin, titin kinase (TK), has been proposed to act as a molecular force sensor in the sarcomer of striated muscles by undergoing force-induced unfolding events presumably linked to its catalytic function. In the first chapter of this thesis, the force-induced unfolding and activation was explored by means of force probe molecular dynamics (FPMD) simulations. In close collaboration with experimental groups, which provided atomic force microscopy (AFM) and enzymatic data, respectively, we showed that this notion of TK being sensitive to external force is indeed correct. A crucial role plays here the removal of autoinhibition, which under normal conditions (i.e. without external force) inhibits catalytic activity of TK.

Previous MD simulations of our group showed that the external force induces the unfolding of the autoregulatory domain, thereby freeing the catalytic site of TK for binding of ATP. The unfolding of the autoregulatory domain left the catalytic site nearly unperturbed, such that catalytic actity was not affected, as required. The forces required to remove autoinhibition were considerably lower than the the forces required to unfold neighbouring titin domains. These results showed that mechanical TK activation might be possible and that TK might indeed play a role as molecular force sensor in muscles.

For the study presented in Chapter 3, kinetic measurements of TK under external AFM-force showed that catalytic activity is gained by exertion of force. Furthermore, in AFM measurements of TK unfolding, saw-tooth shaped force-extension plots were obtained, where the number of peaks corresponded to unfolding events of TK secondary structure elements. In the presence of ATP, one additional force-peak was obtained.

In our FPMD simulations of TK with and without ATP, this additional force peak was also seen in the presence of ATP. We were able to show that this force peak is indeed induced by interactions of the N-terminal TK domain with ATP. After removal of the autoinhibitory tail, the catalytic site of TK is free for ATP to bind. Our simulations correctly predicted that two residues of TK, Met34 and Lys36, were crucial for catalysis. Indeed, in subsequent kinetic measurements, mutations of Lys36 to alanine resulted in almost completely vanished TK catalytic activity.

However, the comparison of simulated force-extension curves with experimentally obtained ones, was not straightforward due to necessarily much higher pulling velocities and much stiffer "cantilevers" in the simulation than in the AFM experiments. Accordingly, we resorted to contour histograms, which display unfolding barriers as a function of

contour lengths which were obtained from a worm-like chain fit to the force-extension curves, thus allowing to quantitatively relate the structural events seen in the simulations to experimentally obtained AFM force-extension plots. The very good agreement of the contour histograms derived from simulations and from experiments, respectively, corroborated that the main unfolding events were described correctly by the simulations.

Overall, the results of this project show that the catalytic activity of TK is indeed force dependent. The much lower mechanical stability of TK as compared to adjacent titin domains enables it to alter its activity by forces which are very well in the range of the physiological forces generated by the actin filaments in sarcomers.

Estimating configurational entropies of macromolecules: the minimally coupled subspace approach

The second objective of this thesis was to find means of quantifying protein dynamics and changes thereof by thermodynamic measures. Chapters 4–6 of this thesis presented a new approach for accurate estimation of configurational entropies of macromolecules. These have to be considered the main driving forces behind many biological processes including molecular recognition, ligand binding, protein folding or other phenomena driven by hydrophobic forces. There is also growing evidence that entropies, and thus dynamics, can be viewed as dynamic carrier of allosteric free energy governing enzymatic activity in finely-tuned complex enzymes. Acccurate entropy calculation from structure ensembles generated by atomistic simulations, however, is not straightforward. For macromolecules of N atoms, the necessary evaluation over the $3N$ dimensional configurational probability density integral has been shown to be nearly impossible. Hence, direct established methods to approximate this integral, such as the widely used (quasi-) harmonic approximation (QH), are limited by neglection of non-linear and higher-order correlations. Non-parametric density estimation allows, in principle, to overcome these limitations; but non-parametric density estimators have so far yielded inaccurate results even for 12-15 degrees of freedom.

In Chapters 4–6, a new hybrid approach to estimate configurational entropies was developed, which incorporates three building blocks. Chapter 4 gives an overview on the concept, and illustrates the applicability to a wide range of macromolecules. The central building block of this method, a non-parametric density estimation method is established (Chapter 5), which is based on adaptive anisotropic kernels. With this

method, non-parametric entropy estimation is shown to be possible of up to 45-dimensional configuration spaces, pushing previous limits by a factor of three. Accuracy was tested and verified for a wide range of dimensionality by comparing our kernel entropy estimates of seven alkanes ranging from butane to decane and the dipeptide dialanine by to two thermodynamic integration (TI) references. It was shown that, in contrast to the QH approximation, adaptive anisotropic kernels reproduced the TI reference with an error of only 4 %. However, for a 14 residue β-turn with more than 500 degrees of freedom, due to the 'curse of dimensionality', which affects all non-parametric density estimation methods, it was impossible to improve on the QH estimate by use of the kernel density estimator alone.

Thus, as the second building block, we employed Full Correlation Analysis (FCA), which uses a linear entropy invariant coordinate transformation such that the usually highly coupled degrees of freedom separate into minimally coupled subspaces each of which being sufficiently small to render non-parametric density estimation applicable. FCA minimizes non-linear correlations of second and higher order [97] and therefore improves on principal component analysis (PCA) which only considers linear correlations of second order. For complex macromolecules, even for the optimal linear transformation, considerable correlations between several degrees of freedom will remain and cannot be neglected. To address this issue, the FCA modes are subsequently clustered according to the generalized correlation coefficient yielding the required sufficiently small subspaces in many cases. This has has been demonstrated exemplary for the coldshock protein and for the apo form of calmodulin (CaM), a large 165-residual calcium regulated allosteric protein.

However, for larger molecules, the above procedure may result in minimally coupled subspaces which are too large for non-parametric density estimation. The third building block of our hybrid approach addresses this issue by subdividing each oversized cluster into a number of smaller subclusters, irrespective of the necessarily remaining strong correlations between these. The correlations which are thus neglected are accounted for via a mutual information expansion (MIE). Previous attempts to apply similar expansions to macromolecules failed due to the "combinatorical explosion" caused by the large number of correlation terms to be considered. We have shown that the clustering of modes achieved via FCA improved the situation considerably, such that the "combinatorical explosions" are avoided.

In combination, these three building blocks enable one to calculate configurational

entropies even for larger biomolecules. As an example of biological relevance, the 146-residue globular protein calmodulin was considered for which scanning or perturbation methods do not converge. Calmodulin (CaM) is a calcium regulated protein which binds up to four calcium ions under physiological conditions. Only the calcium bound structure binds many different peptide binding partners with high affinity. As no direct interaction of calcium with the peptide binding partners occurs, lowering the configurational entropy of the free (but calcium bound) calmodulin has been suggested as a possible regulation mechanism. However, calorimetric experiments suggested a nearly vanishing configurational entropy contribution. We showed that, using the MCSA, a nearly complete entropy cancellation within the error is obtained, in agreement with the experiment. Our result suggested that calcium binding causes a substantial but free-energy-neutral transfer of configurational entropy into solvent entropy, thereby activating CaM for substrate binding. Quite a different picture would be provided by the QH approximation, which clearly overestimated the effect of entropy reduction by a factor of two.

Allosteric regulation of l.mexicana pyruvate kinase

The third part of this thesis aimed at exploring the allosteric regulation of *l.mexicana* pyruvate kinase (PYK), a huge 2000 residue tetrameric enzyme catalysing last step of glycolysis, the conversion of phospho-enol pyruvate to pyruvate. The catalytic activity of PYK is remote-controlled by fructose-bisphosphate, which binds to a binding site more than 2 nm apart (thus the name allostery) from the catalytic site of the same subunit and more than 10 nm to the catalytic binding site of neighbouring subunits. Insight into the allosteric mechanisms of PYK was hitherto limited by the fact that no isoenzyme has crystallised in both active R and inactive T state. Only most recently, a low resolution X-ray structure of a putative *l.mexicana* PYK R-state has been solved, which, due to non-native ligands in both allosteric and catalytic site, was thought not to have assumed a fully activated structure. Our results from extensive MD simulations, which compared the putative *l.mexicana* R-state with the guaranteed fully active *yeast* R-state, in contrast, suggest that the structure has indeed to be considered fully active.

We furthermore showed that, surprisingly for a molecule of this size, pyruvate kinase is regulated on time-scales which are very well accessible to atomistic simulation techniques, despite a millisecond turnover number. Activation and deactivation appear to follow

a classic induced fit mechanism, exhibiting spontaneous T → R (R → T) transition upon introduction (removal) of ligands. The dynamics of the four subunits is largely independent, in disagreement with the symmetry assumption of the MWC model, but in line with the Pauling-KNF model.

We examined the homotropic activation of PYK and observed changes of collective motions throughout the overall tetramer upon binding of ligands at only one or two monomers. These changes of collective motions nicely explain activation events observed in the simulations. The observed transition patterns suggest a dead-end mechanism rather than a hindered first-step mechanism as possible reason why homotropic activation exhibits slower kinetics than heterotropic activation. We proposed kinetic experiments to test this idea.

Concluding remarks

This thesis presented examinations of the homotropic PYK regulation, but not detailed accounts of heterotropic regulation. The examination of heterotropic PYK regulation is limited by the large number of in different intermediate steps, which need to be considered for full description of allosteric activation pathways. With increasing computer performance, however, this issue will certainly be possible to address. A further interesting aspect of PYK allosteric regulation not yet fully investigated is an examination of the role of configurational entropies. The minimally coupled subspace approach developed in Chapters 4–6 now allows to address this issue by proper treatment also of the non-linear and higher order correlations that have been found to be crucial for the correct description of entropy contributions to the regulation of the allosterically regulated calmodulin.

9 Chapter 9
Appendix A

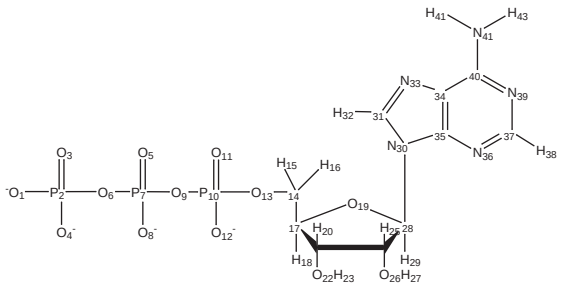

Figure 9.1: Schematic drawing of ATP, defining the atom numbers used in the table.

OPLS type	σ	ϵ
opls_966	0.374	0.837
opls_967	0.350	0.276
opls_968	0.296	0.879
opls_969	0.296	0.879

Table 9.2: Lennard-Jones (6,12) parameters.

#	OPLS type	charge	#	OPLS type	charge	#	OPLS type	charge
1	opls_969	-1.1	16	opls_140	0.06	30	opls_354B	-0.38
2	opls_966	1.9	17	opls_183	0.17	31	opls_353	0.2
3	opls_969	-1.1	18	opls_140	0.03	32	opls_359	0.14
4	opls_969	-1.1	19	opls_180	-0.6	33	opls_352	0.2
5	opls_968	-0.6	20	opls_158	0.2	34	opls_350	0.15
6	opls_969	-1.1	21	opls_140	0.06	35	opls_349	0.38
7	opls_966	1.9	22	opls_171	-0.68	36	opls_348	-0.55
8	opls_969	-1.1	23	opls_172	0.42	37	opls_347	0.35
9	opls_968	-0.7	24	opls_158	0.2	38	opls_355	0.2
10	opls_966	1.9	25	opls_140	0.06	39	opls_346	-0.53
11	opls_969	-1.1	26	opls_171	-0.68	40	opls_351	0.53
12	opls_969	-1.1	27	opls_172	0.42	41	opls_356	-0.74
13	opls_968	-0.7	28	opls_183	0.3	42	opls_357	0.385
14	opls_183	-0.12	29	opls_140	0.06	43	opls_358	0.355
15	opls_140	0.06						

Table 9.4: Atomic partial charges for ATP.

10 Chapter 10
Appendix B

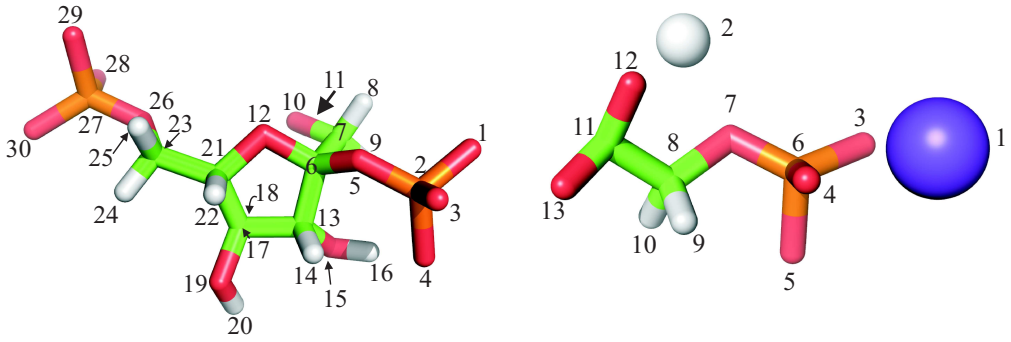

Figure 10.1: Left, structure of fructose-2,6-bisphosphate, defining the atom numbers used in Table 10.2. Right: Structure of PG with K^+ and Mg^{2+}, defining the atom numbers used in Table 10.4

#	OPLS type	charge	#	OPLS type	charge
1	opls_969	-0.983	16	opls_170	0.512
2	opls_966	1.154	17	opls_158	0.187
3	opls_969	-0.905	18	opls_140	0.162
4	opls_969	-0.982	19	opls_169	-0.879
5	opls_968	-0.841	20	opls_170	0.53
6	opls_183	0.694	21	opls_183	0.137
7	opls_157	-0.01	22	opls_140	0.103
8	opls_140	0.18	23	opls_987	0.064
9	opls_140	0.114	24	opls_140	0.075
10	opls_154	-0.827	25	opls_140	0.13
11	opls_170	-0.464	26	opls_968	-0.781
12	opls_180	-0.723	27	opls_966	1.472
13	opls_158	-0.084	28	opls_969	-0.912
14	opls_140	0.167	29	opls_969	-0.962
15	opls_169	-0.882	30	opls_969	-0.947

Table 10.2: Atomic partial charges for FBP

#	OPLS type	charge	#	OPLS type	charge
1	opls_411	1.295	8	opls_967	-0.099
2	opls_408	0.86	9	opls_140	0.256
3	opls_969	-0.843	10	opls_140	0.258
4	opls_966	1.395	11	opls_235	0.742
5	opls_969	-0.881	12	opls_236	-0.687
6	opls_969	-0.866	13	opls_236	-0.684
7	opls_968	-0.773			

Table 10.4: Atomic partial charges for PG with K^+ and Mg^{2+}

11 Chapter 11

Acknowledgements

I would like to thank my supervisor Helmut Grubmüller, for providing me with the opportunity to study for the degree of Doctor of natural things, and for his continuous advice and encouragement throughout the duration of my study; and, to be exact, even before when he kindly supported my idea to carry out the first part of my study at the University of Edinburgh as a visitor in Malcolm D. Walkinshaw's structural biology group and let choose the first project of my thesis on my own. Surely not many allow their students such big freedom to follow their own ideas, whilst also providing them with the right impulses at the right time. He always had an open ear for any questions related to all the projects presented in this thesis that came up in Göttingen and when I was in Scotland. Not least, he managed to create an open and positive atmosphere in all discussions, such that even critical comments were instructive and motivating rather than disencouraging. And he only wanted four beers for going through this thesis superquickly.

I am very grateful for the friendship and all the help of Oliver Lange connected and not connected with this thesis. He supervised the entropy part of this thesis on a daily basis, and certainly this thesis would be thinner by exactly three chapters without his nearly unlimited reservoir of good ideas at the right time and his refusal to believe that it might not work. He also made numerous helpful suggestions to the allostery project. Also I would most likely not have discovered that Tesco's actually does sell reasonable beer – by Scottish measures, and that Göttingen is a good place to live despite the weather.

Many thanks also to Martin Suhm for acting as the first referee of this thesis, for accepting me as a funded member of the physical chemistry graduate school and for the possibility to attend three graduate school workshops in Germany and France, one of which I was given the opportunity to organise. The graduate school also funded my

attendance of three most stimulating conferences in Europe and the US.

I thank Prof. Malcolm Walkinshaw, who was so kind to accept me as a guest researcher in his structural biology group and not to call me insane when he learnt that I wanted to simulate a 60 kDa enzyme in explicit solvent rather than improve docking algorithms. During the first 1.5 years of my thesis, he was the supervisor in charge on a daily basis and was always ready to help. Due to the weekly group meetings I now almost understand what structural biology is all about. I acknowledge the warm welcome and support by his group, in particular Paul Taylor, Conny Ludwig, Kirsten Lillie, and the 'Taiwanese lab'.

I very much thank the members of the ivory tower, the Theoretical and Computational Biophysics Department of the MPI for Biophysical Chemistry for a positive and creative atmosphere, useful discussions, helpful suggestions and the like, and all the fun. Of all of the nice members of this nice group, I'd like to mention a few which had particular impact to this thesis, in particular Lars V. Schäfer, who has a weird italo-saxonian accent but is a good man and a clever scientist. And a friend, who made unfolding titin kinase a thing to look forward to in the morning. He, together with Oliver Lange, Matthias Müller and Bert de Groot, also helped very much during my start-up phase of the PhD study answering questions concerning gromacs. Bert also had a considerable number of most helpful suggestions concerning almost all parts of this thesis. Thanks to Marcus Hennig, who made me assume a whole new view on mathematics, office naps and pasta for breakfast, lunch and dinner; and introduced me to a number of most interesting views and tools. Thanks also to Frauke Gräter, Frank Wiederschein, Jürgen Haas and Ulrich Zachariae, who were always there to distract me from work.

I also thank Elias Puchner, Hermann Gaub and Mathias Gautel for the most productive collaboration related to the titin kinase project.

Thanks to Ansgar Eszterman, Martin Fechner and Evi Heinemann, who always provided superb working environment and great administrative support.

This thesis would not exist without the help of my family and friends. Thanks ever so much for the support.

Bibliography

[1] A. M. Lesk. *Introduction to Protein Architecture*. Oxford University Press,Oxford, 2001.

[2] B. F. Rasmussen, A. M. Stock, D. Ringe, and G. A. Petsko. Crystalline ribonuclease A loses function below the dynamical transition at 220 K. *Nature*, 357(6377):423–424, June 1992.

[3] W. Hoppe, W. Lohmann, H. Markl, and H. Ziegler, editors. *Biophysik*. Springer, 1982.

[4] J. W. Jung and W. Lee. Structure-based functional discovery of proteins: Structural proteomics. *J. Biochem. Mol. Biol.*, 37:28–34, 2004.

[5] A. T. Brunger and M. Nilges. Computational challenges for macromolecular structure determination by Xray crystallography and solution NMR spectroscopy. *Q. Rev. Biophys.*, 26:49–125, 1993.

[6] M. Nilges. Structure calculation from NMR data. *Curr. Opin. Struct. Biol.*, 6:617–623, 1996.

[7] V. Srajer, T. Y. Teng, T. Ursby, C. Pradervand, Z. Ren, S. Adachi, W. Schildkamp, D. Bourgeois, M. Wulff, and K. Moffat. Photolysis of the carbon monoxide complex of myoglobin: Nanosecond time-resolved crystallography. *Science*, 274(5293):1726–1729, 1996.

[8] V. Srajer, Z. Ren, T. Y. Teng, M. Schmidt, T. Ursby, D. Bourgeois, C. Pradervand, W. Schildkamp, M. Wulff, and K. Moffat. Protein conformational relaxation and ligand migration in myoglobin: A nanosecond to millisecond molecular movie from time-resolved Laue X-ray diffraction. *Biochemistry*, 40(46):13802–13815, 2001.

[9] E. Z. Eisenmesser, D. A. Bosco, M. Akke, and D. Kern. Enzyme dynamics during catalysis. *Science*, 295(5559):1520–1523, 2002.

[10] R. Brüschweiler. New approaches to the dynamic interpretation and prediction of NMR relaxation data from proteins. *Curr. Opin. Struct. Biol.*, 13(2):175–183, 2003.

[11] J. G. Kempf and J. Loria. Protein dynamics from solution NMR theory and applications. *Cell Biochem. Biophys.*, 37:187–211, 2003.

[12] J. C. Smith. Protein dynamics — comparison of simulations with inelastic neutron-scattering experiments. *Q. Rev. Biophys.*, 24:227–291, 1991.

[13] F. Gabel, D. Bicout, U. Lehnert, M. Tehei, M. Weik, and G. Zaccai. Protein dynamics studied by neutron scattering. *Q. Rev. Biophys.*, 35:327–367, 2002.

[14] F. Garczarek and K. Gerwert. Functional waters in intraprotein proton transfer monitored by FTIR difference spectroscopy. *Nature*, 439(7072):109–112, 2006.

[15] E. T. J. Nibbering, H. Fidder, and E. Pines. Ultrafast chemistry: Using time-resolved vibrational spectroscopy for interrogation of structural dynamics. *Annu. Rev. Phys. Chem.*, 56:337–367, 2005.

[16] C. Kotting and K. Gerwert. Proteins in action monitored by time-resolved FTIR spectroscopy. *ChemPhysChem*, 6(5):881–888, 2005.

[17] H. Dietz and M. Rief. Protein structure by mechanical triangulation. *Proc. Natl. Acad. Sci. U. S. A.*, 103(5):1244–1247, 2006.

[18] R. H. Zhou, E. Harder, H. F. Xu, and B. J. Berne. Efficient multiple time step method for use with Ewald and particle mesh Ewald for large biomolecular systems. *J. Chem. Phys.*, 115(5):2348–2358, 2001.

[19] M. E. Tuckerman, B. J. Berne, and G. J. Martyna. Molecular-dynamics algorithm for multiple time scales - systems with long-range forces. *J. Chem. Phys.*, 94(10):6811–6815, 1991.

[20] M. Tuckerman, B. J. Berne, and G. J. Martyna. Reversible multiple time scale molecular-dynamics. *J. Chem. Phys.*, 97(3):1990–2001, 1992.

[21] P. Minary, M. E. Tuckerman, and G. J. Martyna. Long time molecular dynamics for enhanced conformational sampling in biomolecular systems. *Phys. Rev. Lett.*, 93(15):150201, 2004.

[22] J. L. Scully and J. Hermans. Multiple time steps - limits on the speedup of molecular-dynamics simulations of aqueous systems. *Mol. Simul.*, 11(1):67–77, 1993.

[23] F. Zhang. Operator-splitting integrators for constant-temperature molecular dynamics. *J. Chem. Phys.*, 106(14):6102–6106, 1997.

[24] J. A. Board, J. W. Csey, J. F. Leathrum, A. Windemuth, and K. Schulten. Accelerated molecular-dynamics simulation with the parallel fast multipole algorithm. 198(1-2):89–94, 1992.

[25] A. M. Mathiowetz, A. Jain, N. Karasawa, and W. A. Goddard. Protein simulations using techniques suitable for very large systems - the cell multipole method for nonbond interactions and the Newton-Euler inverse mass operator method for internal coordinate dynamics. *Proteins*, 20(3):227–247, 1994.

[26] M. Eichinger, H. Grubmüller, H. Heller, and P. Tavan. FAMUSAMM: An algorithm for rapid evaluation of electrostatic interactions in molecular dynamics simulations. *J. Comput. Chem.*, 18(14):1729–1749, 1997.

[27] L. Greengard and V. Rokhlin. On the evaluation of electrostatic interactions in molecular modeling. 29A:139–144, 1989.

[28] A. Y. Toukmaji and J. A. Board. Ewald summation techniques in perspective: A survey. *Comput. Phys. Commun.*, 95(2-3):73–92, 1996.

[29] J. P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. Numerical-integration of cartesian equations of motion of a system with constraints - molecular-dynamics of n-alkanes. *J. Comput. Phys.*, 23(3):327–341, 1977.

[30] S. Miyamoto and P. A. Kollman. Settle: An analytical version of the shake and rattle algorithm for rigid water models. *J. Comp. Chem.*, 13(8):952–962, 1992.

[31] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije. Lincs: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry*, 18(12):1463–1472, 1997.

[32] K. Tai. Conformational sampling for the impatient. *Biophysical Chemistry*, 107(3):213–220, February 2004.

[33] H. Grubmüller, B. Heymann, and P. Tavan. Ligand binding: Molecular mechanics calculation of the streptavidin-biotin rupture force. *Science*, 271(5251):997–999, 1996.

[34] V. T. Moy, E. L. Florin, and H. E. Gaub. Intermolecular forces and energies between ligands and receptors. *Science*, 266(5183):257–259, 1994.

[35] M. Rief, M. Gautel, F. Oesterhelt, J. M. Fernandez, and H. E. Gaub. Reversible unfolding of individual titin immunoglobulin domains by afm. *Science*, 276(5315):1109–1112, 1997.

[36] T. E. Fisher, P. E. Marszalek, and J. M. Fernandez. Stretching single molecules into novel conformations using the atomic force microscope. *Nat Struct Mol Biol*, 7(9):719–724, September 2000.

[37] J. A. Hardy and J. A. Wells. Searching for new allosteric sites in enzymes. *Current Opinion in Structural Biology*, 14(6):706–715, December 2004.

[38] Z.-G. Gao and K. A. Jacobson. Keynote review: Allosterism in membrane receptors. *Drug Discovery Today*, 11(5-6):191–202, March 2006.

[39] J. F. Swain and L. M. Gierasch. The changing landscape of protein allostery. *Current Opinion in Structural Biology*, 16(1):102–108, February 2006.

[40] G. M. Suel, S. W. Lockless, M. A. Wall, and R. Ranganathan. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat Struct Mol Biol*, 10(1):59–69, January 2003.

[41] N. Popovych, S. Sun, R. H. Ebright, and C. G. Kalodimos. Dynamically driven protein allostery. *Nat Struct Mol Biol*, 13(9):831–838, September 2006.

[42] J.-P. Changeux and S. J. Edelstein. Allosteric mechanisms of signal transduction. *Science*, 308(5727):1424–1428, 2005.

[43] T. A. J. Duke, N. Le Novère, and D. Bray. Conformational spread in a ring of proteins: A stochastic approach to allostery. *Journal of Molecular Biology*, 308(3):541–553, May 2001.

[44] A. Cooper and D. T. F. Dryden. Allostery without conformational change. *European Biophysics Journal*, 11(2):103–109, October 1984.

[45] L. Li, V. N. Uversky, A. K. Dunker, and S. O. Merueh. A computational investigation of allostery in the catabolic activator protein. *J. Am. Chem. Soc.*, 129:15668–15676, 2007.

[46] S. Y. Stevens, S. Sanker, C. Kent, and E. R. Zuiderweg. Delineation of the allosteric mechanism of a cytidylyltransferase exhibiting negative cooperativity. *Nat Struct Mol Biol*, 8(11):947–952, November 2001.

[47] D. E. Koshland. The structural basis of negative cooperativity: receptors and enzymes. *Current Opinion in Structural Biology*, 6(6):757–761, December 1996.

[48] L. Tskhovrebova and J. Trinick. Titin: properties and family relationships. *Nat Rev Mol Cell Biol*, 4(9):679–689, September 2003.

[49] M. Hoshijima. Mechanical stress-strain sensors embedded in cardiac cytoskeleton: Z disk, titin, and associated structures. *Am J Physiol Heart Circ Physiol*, 290(4):H1313–1325, 2006.

[50] F. Gräter, J. Shen, H. Jiang, M. Gautel, and H. Grubmüller. Mechanically induced titin kinase activation studied by force-probe molecular dynamics simulations. *Biophys. J.*, 88(2):790–804, 2005.

[51] D. N. Greene, T. Garcia, R. B. Sutton, K. M. Gernert, G. M. Benian, and A. F. Oberhauser. Single-molecule force spectroscopy reveals a stepwise unfolding of caenorhabditis elegans giant protein kinase domains. *Biophys. J.*, 95(3):1360–1370, 2008.

[52] E. M. Puchner, G. Franzen, M. Gautel, and H. E. Gaub. Comparing proteins by their unfolding pattern. *Biophys. J.*, 95(1):426–434, 2008.

[53] D. L. Beveridge and F. M. DiCapua. Free energy via molecular simulation: Applications to chemical and biomolecular systems. *Annual Review of Biophysics and Biophysical Chemistry*, 18(1):431–492, 1989.

[54] P. Kollman. Free energy calculations: Applications to chemical and biochemical phenomena. *Chem. Rev.*, 93(7):2395–2417, 1993.

[55] M. K. Gilson and H.-X. Zhou. Calculation of protein-ligand binding affinities. *Ann. Rev. Biophys. Biomol. Struct.*, 36(1):21–42, 2007.

[56] J. N. Karplus, Martin; Kushick. Method for estimating the configurational entropy of macromolecules. *Macromolecules*, 14(2):325–332, 1981.

[57] J. Schlitter. Estimation of absolute and relative entropies of macromolecules using the covariance matrix. *Chemical Physics Letters*, 215(6):617–621, December 1993.

[58] W. F. van Gunsteren, D. Bakowies, R. Baron, I. Chandrasekhar, M. Christen, X. Daura, P. Gee, D. P. Geerke, A. Glättli, P. H. Hünenberger, M. A. Kastenholz, C. Oostenbrink, M. Schenk, D. Trzesniak, N. F. A. van der Vegt, and H. B. Yu. Biomolecular modeling: Goals, problems, perspectives. *Angewandte Chemie International Edition*, 45(25):4064–4092, 2006.

[59] A. Leach. *Molecular Modelling: Principles and Applications*. Pearson Education Ltd., Essex, England, 2001.

[60] F. Jensen. *Computational Chemistry*. John Wiley and Sons Ltd., Sussex, England, 2001.

[61] M. Born and R. Oppenheimer. Zur Quantentheorie der Molekeln. *Ann. d. Physik*, 84(20):457–484, 1927.

[62] R. W. Hockney. The potential calculation and some applications. *Methods Comput. Phys.*, 9:136–211, 1970.

[63] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79:926–935, 1983.

[64] G. Kaminski, R. Friesner, J. Tirado-Rives, and W. Jorgensen. Evaluation and reparametrization of the opls-aa force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B*, 105(28):6474–6487, 2001.

[65] W. F. van Gunsteren, X. Daura, and A. E. Mark. *GROMOS force field*, pages 1211–1216. Encyclopaedia of computational chemistry edition, 1998.

[66] J. M. Wang, P. Cieplak, and P. A. Kollman. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.*, 21(12):1049–1074, 2000.

[67] A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, 102(18):3586–3616, 1998.

[68] N. L. Allinger, Y. H. Yuh, and J.-H. Lii. Molecular mechanics. The MM3 force fields for hydrocarbons. 111:8551–8566, 1989.

[69] 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(8):3684–3690, October 1984.

[70] M. Saito. Molecular dynamics of proteins in solution – artifacts by the cutoff approximation. *J. Phys. Chem.*, 101:4055–4061, 1994.

[71] P. Ewald. Berechnung optischer und elektrostatischer Gitterpotentiale. *Ann. d. Physik*, IV:253–287, 1921.

[72] 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(12):10089–10092, June 1993.

[73] Picture courtesy of E. Puchner.

[74] C. Bouchiat, M. D. Wang, J. F. Allemand, T. Strick, S. M. Block, and V. Croquette. Estimating the persistence length of a worm-like chain molecule from force-extension measurements. *Biophys. J.*, 76(1):409–413, 1999.

[75] H. Grubmüller, B. Heymann, and P. Tavan. Ligand binding: Molecular mechanics calculation of the streptavidin-biotin rupture force. *Science*, 271(5251):997–999, 1996.

[76] H. Grubmüller. *Protein-Ligand Interactions*. Humana Press Inc., Towata, NJ, 2005.

[77] B. Heymann and H. Grubmüller. Chair-boat transitions and side groups affect the stiffness of polysaccharides. *Chem. Phys. Lett.*, 305(3-4):202–208, 1999.

[78] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572, 1901.

[79] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience, 2 edition, 2000.

[80] K. Karhunen. über lineare methoden in der wahrscheinlichkeitsrechnung. *Ann. Acad. Sci. Fennicae. Ser. A. I. Math.-Phys.*, 37:1–79, 1947.

[81] R. M. Levy, M. Karplus, J. Kushick, and D. Perahia. Evaluation of the configurational entropy for proteins - application to molecular-dynamics simulations of an alpha-helix. *Macromolecules*, 17(7):1370–1374, 1984.

[82] R. M. Levy, A. R. Srinivasan, W. K. Olson, and J. A. McCammon. Quasi-harmonic method for studying very low-frequency modes in proteins. *Biopolymers*, 23(6):1099–1112, 1984.

[83] M. M. Teeter and D. A. Case. Harmonic and quasiharmonic descriptions of crambin. *J. Phys. Chem.*, 94(21):8091–8097, 1990.

[84] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika I*, pages 211–218, 1936.

[85] M. E. Wall, A. Rechtsteiner, and L. M. Rocha. *A Practical Approach to Microarray Data Analysis*, chapter Singular value decomposition and principal component analysis, pages 91–109. Kluwer: Norwell, MA, 2003.

[86] I. Bahar, B. Erman, T. Haliloglu, and R. L. Jernigan. Efficient characterization of collective motions and interresidue correlations in proteins by low-resolution simulations. *Biochemistry*, 36(44):13512–13523, 1997.

[87] T. D. Romo, J. B. Clarage, D. C. Sorensen, and G. N. Phillips. Automatic identification of discrete substates in proteins - singular-value decomposition analysis of time-averaged crystallographic refinements. *Proteins*, 22(4):311–321, 1995.

[88] A. Amadei, A. Linssen, and H. Berendsen. Essential dynamics of proteins. *Proteins*, 17:412–425, 1993.

[89] S. Hayward and N. Go. Collective variable description of native protein dynamics. *Annu. Rev. Phys. Chem*, 46:223–250, 1995.

[90] R. Abseher and M. Nilges. Efficient sampling in collective coordinate space. *Proteins*, 39(1):82–88, 2000.

[91] A. Amadei, A. B. M. Linssen, B. L. de Groot, D. M. F. van Aalten, and H. J. C. Berendsen. An efficient method for sampling the essential subspace of proteins. *J. Biomol. Struct. Dyn.*, 13(4):615–625, 1996.

[92] D. Mustard and D. W. Ritchie. Docking essential dynamics eigenstructures. *Proteins*, 60(2):269–274, 2005.

[93] A. L. Tournier and J. C. Smith. Principal components of the protein dynamical transition. *Phys. Rev. Lett.*, 91(20):208106–1 – 208106–4, 2003.

[94] B. L. de Groot, X. Daura, A. E. Mark, and H. Grubmüller. Essential dynamics of reversible peptide folding: Memory-free conformational dynamics governed by internal hydrogen bonds. *J. Mol. Biol.*, 309(1):299–313, 2001.

[95] P. Billingsley. *Probability and Measure*. 3 edition, 1995.

[96] S. Hayward, A. Kitao, and N. Go. Harmonicity and anharmonicity in protein dynamics — a normal-mode analysis and principal component analysis. *Proteins*, 23(2):177–186, 1995.

[97] O. F. Lange and H. Grubmüller. Full correlation analysis of conformational protein dynamics. *Proteins*, 70(4):1294–1312, 2008.

[98] A. M. Fraser and H. L. Swinney. Independent coordinates for strange attractors from mutual information. *Phys. Rev. A*, 33(2):1134–1140, 1986.

[99] A. Hyvärinen. *Advances in Neural Information Processing Systems*, volume 10, chapter New approximations of differential entropy for independent component analysis and projection pursuit, pages 273–279. MIT Press, 1998.

[100] A. Hyvärinen and P. Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Netw.*, 12(3):429–439, 1999.

[101] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Netw.*, 10(3):626–634, 1999.

[102] A. Hyvärinen and E. Oja. Independent component analysis: Algorithms and applications. *Neural Netw.*, 13(4-5):411–430, 2000.

[103] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley, New York, 2001.

[104] A. Hyvärinen. An alternative approach to infomax and independent component analysis. *Neurocomputing*, 44:1089–1097, 2002.

[105] J. Trinick and L. Tskhovrebova. Titin: a molecular control freak. *Trends in Cell Biology*, 9(10):377–380, October 1999.

[106] M. Kellermayer and L. Grama. Stretching and visualizing titin molecules: combining structure, dynamics and mechanics. *Journal of Muscle Research and Cell Motility*, 23(5):499–511, October 2002.

[107] W. Linke and J. Fernandez. Cardiac titin: molecular basis of elasticity and cellular contribution to elastic and viscous stiffness components in myocardium. *Journal of Muscle Research and Cell Motility*, 23(5):483–497, October 2002.

[108] http://www.othyr.com/titin.html.

[109] O. Mayans, P. F. M. van der Ven, M. Wilm, A. Mues, P. Young, D. O. Furst, M. Wilmanns, and M. Gautel. Structural basis for activation of the titin kinase domain during myofibrillogenesis. *Nature*, 395(6705):863–869, October 1998.

[110] I. Agarkova and J.-C. Perriard. The m-band: an elastic web that crosslinks thick filaments in the center of the sarcomere. *Trends in Cell Biology*, 15(9):477–485, September 2005.

[111] S. Lange, F. Xiang, A. Yakovenko, A. Vihola, P. Hackman, E. Rostkova, J. Kristensen, B. Brandmeier, G. Franzen, B. Hedberg, L. G. Gunnarsson, S. M. Hughes, S. Marchand, T. Sejersen, I. Richard, L. Edstrom, E. Ehler, B. Udd, and M. Gautel. The kinase domain of titin controls muscle gene expression and protein turnover. *Science*, 308(5728):1599–1603, 2005.

[112] E. M. Puchner, A. Alexandrovich, A. L. Kho, U. Hensen, L. V. Schäfer, B. Brandmeier, F. Gräter, H. Grubmüller, H. E. Gaub, and M. Gautel. Mechanoenzymatics of titin kinase. *Proc. Natl. Acad. Sci. U. S. A.*, 105(36):13385–13390, 2008.

[113] D. V. D. Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen. Gromacs: Fast, flexible, and free. *J. Comp. Chem.*, 26(16):1701–1718, 2005.

[114] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl. Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theor. Comp.*, 4:0, 2008.

[115] C. M. Breneman and K. B. Wiberg. Determining atom-centered monopoles from molecular electrostatic potentials. the need for high sampling density in formamide conformational analysis. *Journal of Computational Chemistry*, 11(3):361–373, 1990.

[116] M. J. Frisch *et al.* Gaussian 03, Revision B.04.

[117] M. Klähn, J. Schlitter, and K. Gerwert. Theoretical IR spectroscopy based on QM/MM calculations provides changes in charge distribution, bond lengths, and bond angles of the GTP ligand induced by the Ras-protein. *Biophys. J.*, 88(6):3829–3844, 2005.

[118] M. Gäsel, C. B. Breitenlechner, P. Rüger, U. Jucknischke, T. Schneider, R. Huber, D. Bossemeyer, and R. A. Engh. Mutants of protein kinase A that mimic the ATP-binding site of protein kinase B (AKT). *Journal of Molecular Biology*, 329(5):1021–1034, June 2003.

[119] E. D. Lowe, M. E. Noble, V. T. Skamnaki, N. G. Oikonomakos, D. J. Owen, and J. L. N. The crystal structure of a phosphorylase kinase peptide substrate complex: Kinase substrate recognition. *EMBO J*, 16:6646–6658, 1997.

[120] B. L. de Groot, G. Vriend, and H. J. C. Berendsen. Conformational changes in the chaperonin groel: new insights into the allosteric mechanism. *Journal of Molecular Biology*, 286(4):1241–1249, March 1999.

[121] S. Izrailev, S. Stepaniants, M. Balsera, Y. Oono, and K. Schulten. Molecular dynamics study of unbinding of the avidin-biotin complex. *Biophys. J.*, 72(4):1568–1581, 1997.

[122] T. P. Straatsma and J. A. McCammon. Computational alchemy. *Annual Review of Physical Chemistry*, 43(1):407–435, 1992.

[123] H. Meirovitch. Recent developments in methodologies for calculating the entropy and free energy of biological systems by computer simulation. *Curr. Opin. Struct. Biol.*, 17(2):181–186, April 2007.

[124] C. Peter, C. Oostenbrink, A. van Dorp, and W. F. van Gunsteren. Estimating entropies from molecular dynamics simulations. *J. Chem. Phys.*, 120(6):2652–2661, 2004.

[125] S. Cheluvaraja and H. Meirovitch. Simulation method for calculating the entropy and free energy of peptides and proteins. *Proc. Natl. Acad. Sci. U. S. A.*, 101(25):9241–9246, 2004.

[126] S. Cheluvaraja and H. Meirovitch. Calculation of the entropy and free energy of peptides by molecular dynamics simulations using the hypothetical scanning molecular dynamics method. *J. Chem. Phys.*, 125(2):024905, 2006.

[127] M. Karplus and J. A. McCammon. Molecular dynamics simulations of biomolecules. *Nat Struct Mol Biol*, 9(9):646–652, September 2002.

[128] H. Schäfer, A. E. Mark, and W. F. van Gunsteren. Absolute entropies from molecular dynamics simulation trajectories. *J. Chem. Phys.*, 113(18):7809–7817, November 2000.

[129] C. Chang, W. Chen, and M. Gilson. Evaluating the accuracy of the quasiharmonic approximation. *J. Chem. Theory Comput.*, 1(5):1017–1028, 2005.

[130] M. Tyka, A. Clarke, and R. Sessions. An efficient, path-independent method for free-energy calculations. *J. Phys. Chem. B*, 110(34):17212–17220, 2006.

[131] V. Hnizdo, E. Darian, A. Fedorowicz, E. Demchuk, S. Li, and H. Singh. Nearest-neighbor nonparametric method for estimating the configurational entropy of complex molecules. *J. Comp. Chem.*, 28(3):655–668, 2007.

[132] R. E. Bellman. *Adaptive Control Processes*. Princeton University Press, 1961.

[133] M. Hennig. Entropy invariant transformations. Master's thesis, Universität Jena, 2007.

[134] A. Baranyai and D. J. Evans. Direct entropy calculation from computer simulation of liquids. *Phys. Rev. A*, 40(7):3817–3822, October 1989.

[135] P. Attard, O. G. Jepps, and S. Marčelja. Information content of signals using correlation function expansions of the entropy. *Phys. Rev. E*, 56(4):4052–4067, October 1997.

[136] B. J. Killian, J. Y. Kravitz, and M. K. Gilson. Extraction of configurational entropy from molecular simulations via an expansion approximation. *J. Chem. Phys.*, 127(2):024107, 2007.

[137] V. Hnizdo, J. Tan, B. J. Killian, and M. K. Gilson. Efficient calculation of configurational entropy from molecular simulations by combining the mutual-information expansion and nearest-neighbor methods. *J. Comp. Chem.*, 29(10):1605–1614, 2008.

[138] M. A. Wilson and A. T. Brunger. The 1.0 Å crystal structure of Ca^{2+}-bound calmodulin: An analysis of disorder and implications for functionally relevant plasticity. *J. Mol. Biol.*, 301(5):1237–1256, September 2000.

[139] H. Kuboniwa, N. Tjandra, S. Grzesiek, H. Ren, C. B. Klee, and A. Bax. Solution structure of calcium-free calmodulin. *Nat Struct Mol Biol*, 2(9):768–776, September 1995.

[140] R. Gilli, D. Lafitte, C. Lopez, M.-C. Kilhoffer, A. Makarov, C. Briand, and J. Haiech. Thermodynamic analysis of calcium and magnesium binding to calmodulin. *Biochemistry*, 37(16):5450–5456, 1998.

[141] P. L. Privalov and S. J. Gill. *Stability of Protein Structure and Hydrophobic Interaction*, volume 39, pages 191–234. 1988.

[142] R. Grünberg, M. Nilges, and J. Leckner. Flexibility and conformational entropy in protein-protein binding. *Structure*, 14(4):683–693, April 2006.

[143] H. Gohlke and D. A. Case. Converging free energy estimates: Mm-pb(gb)sa studies on the protein-protein complex ras-raf. *J. Comp. Chem.*, 25(2):238–250, 2004.

[144] J. Haas, E. Vöhringer-Martinez, A. Bögehold, D. Matthes, U. Hensen, A. Pelah, B. Abel, and H. Grubmüller. Primary steps of ph-dependent insulin aggregation kinetics are governed by conformational flexibility. *ChemBioChem*, accepted.

[145] N. Go and H. A. Scheraga. Analysis of the contribution of internal vibrations to the statistical weights of equilibrium conformations of macromolecules. *J. Chem. Phys.*, 51:4751, 1969.

[146] D. M. F. Schüttelkopf, A. W.; van Aalten. Prodrg - a tool for high-throughput crystallography of protein-ligand complexes. *Acta Crystallographica D*, 60:1355–1363, 2004.

[147] http://www.cs.umd.edu/∼mount/ann/.

[148] J. G. Kirkwood. Statistical mechanics of fluid mixtures. *The Journal of Chemical Physics*, 3(5):300–313, 1935.

[149] L. Comtet. *Advanced Combinatorics: The Art of Finite and Infinite Expansions*, volume 1. Reidel, Dordrecht/Boston, 1974.

[150] R. Yeung. A new outlook on shannon's information measures. *Information Theory, IEEE Transactions on*, 37(3):466–474, 1991.

[151] T. Kawabata and R. Yeung. The structure of the I-measure of a markov chain. *Information Theory, IEEE Transactions on*, 38(3):1146–1149, 1992.

[152] E. T. Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106(4):620–630, May 1957.

[153] J. G. Kirkwood and E. M. Boggs. The radial distribution function in liquids. *The Journal of Chemical Physics*, 10(6):394–402, 1942.

[154] I. Z. Fisher and B. Kopeliovich. *Proc. Acad. Sci. USSR*, 133:81, 1960.

[155] S. Watanabe. Information theoretical analysis of mulitvariate correlation. *IBM Journal*, 4:66–82, January 1960.

[156] P. Attard. *Statistical Physics on the Eve of the Twenty-First Century*, chapter Markov Superposition Expansion for the Entropy and Correlation Functions in Two and Three Dimensions. World Scientific, 1999.

[157] J. Numata. Conformational entropy from biomolecular simulation using information theory. In *Contribution to conference: Entropy in Biomolecular Systems, Split*, 2008.

[158] H. Matsuda. Physical nature of higher-order mutual information: Intrinsic correlations and frustration. *Phys. Rev. E*, 62(3):3096–3102, September 2000.

[159] H. Frauenfelder and P. G. Wolynes. *Phys. Today*, 47:58–64, 1994.

[160] L. Pauling. The oxygen equilibrium of hemoglobin and its structural interpretation. *Proc. Natl. Acad. Sci.*, 21:181–191, 1935.

[161] J. Monod and F. Jacob. Teleonomic mechanisms in cellular metabolism, growth and differentiation. *Cold Spring Harb. Symp. Quant. Biol.*, 26:389–401, 1961.

[162] D. E. Koshland, G. Nemethy, and D. Filmer. Comparison of experimental binding data and theoretical models in proteins containing subunits. *Biochemistry*, 5:365–385, 1966.

[163] J. Monod, J. Wyman, and J.-P. Changeux. On the nature of allosteric transitions: a plausible model. *J. Mol. Biol.*, 12:88–118, 1965.

[164] A. Hill. The possible effects of the aggregation of the molecules of haemoglobin on its oxygen dissociation. *J. Physiol.*, 40:iv–vii, 1910.

[165] J. Weiss. The hill equation revisited: uses and misuses. *FASEB J.*, 11(11):835–841, 1997.

[166] M. Perutz. X-ray analysis of haemoglobin. *Science*, 140:863–869, 1963.

[167] M. Perutz, W. Bolton, R. Diamond, H. Muirhead, and H. Watson. Structure of haemoglobin: an x-ray examination of reduced horse haemoglobin. *Nature*, 203:687–690, 1964.

[168] J.-P. Changeux and S. J. Edelstein. Allosteric receptors after 30 years. *Neuron*, 21(5):959–980, November 1998.

[169] A. Szabo and M. Karplus. A mathematical model for structure-function relations in hemoglobin. *Journal of Molecular Biology*, 72(1):163–197, December 1972.

[170] A. Ansari, H. J. Berendsen, S. F. Bowne, H. Frauenfelder, I. E. Iben, T. B. Sauke, E. Shyamsunder, and R. D. Young. Protein states and proteinquakes. *Proc. Natl. Acad. Sci. U. S. A.*, 82(15):5000–5004, 1985.

[171] C. H. Robert, H. Decker, B. Richey, S. J. Gill, and J. Wyman. Nesting: Hierarchies of allosteric interactions. *Proc. Natl. Acad. Sci. U. S. A.*, 84(7):1891–1895, 1987.

[172] G. Ackers, M. Doyle, D. Myers, and M. Daugherty. Molecular code for cooperativity in hemoglobin. *Science*, 255:54–63, 1992.

[173] Y. Huang and G. Ackers. Enthalpic and entropic components of cooperativity for the partially ligated intermediates of hemoglobin support a "symmetry rule" mechanism. *Biochemistry*, 34:6316–6327, 1995.

[174] V. L. Tlapak-Simmons and G. D. Reinhart. Obfuscation of allosteric structure-function relationships by enthalpy-entropy compensation. *Biophys. J.*, 75(2):1010–1015, 1998.

[175] E. R. Henry, S. Bettati, J. Hofrichter, and W. A. Eaton. A tertiary two-state allosteric model for hemoglobin. *Biophysical Chemistry*, 98(1-2):149–164, July 2002.

[176] Stryer. *Biochemistry*. Freeman and Co., San Francisco, 1975.

[177] W. Tian, T. Sage, P. Champion, E. Chien, and S. Sligar. Probing heme protein conformational equilibration rates with kinetic selection. *Biochemistry*, 35:3487–3502, 1996.

[178] E. Freire. The propagation of binding interactions to remote sites in proteins: Analysis of the binding of the monoclonal antibody d1.3 to lysozyme. *Proc. Natl. Acad. Sci. U. S. A.*, 96(18):10118–10122, 1999.

[179] E. W. Yu and D. E. Koshland. Propagating conformational changes over long (and short) distances in proteins. *Proc. Natl. Acad. Sci. U. S. A.*, 98(17):9517–9520, 2001.

[180] B. Jacrot, S. Cusack, A. Dianoux, and D. Engelman. Inelastic neutron scattering analysis of hexokinase dynamics and its modification of binding of glucose. *Nature*, 300:84–86, 1982.

[181] H. Middendorf. Biophysical applications of quasi-elastic and inelastic neutron scattering. *Annual Review of Biophysics and Bioengineering*, 13:425–451, 1984.

[182] G. Weber. Ligand binding and internal equilibria in proteins. *Biochemistry*, 11:864–878, 1972.

[183] K. Gunasekaran, B. Ma, and R. Nussinov. Is allostery an intrinsic property of all dynamic proteins? *Proteins: Structure, Function, and Bioinformatics*, 57(3):433–443, 2004.

[184] K. P. Ravindranathan, E. Gallicchio, and R. M. Levy. Conformational equilibria and free energy profiles for the allosteric transition of the ribose-binding protein. *Journal of Molecular Biology*, 353(1):196–210, October 2005.

[185] D. Kern and E. R. Zuiderweg. The role of dynamics in allosteric regulation. *Curr. Opin. Struct. Biol.*, 13:748–757, 2003.

[186] I. Luque, S. A. Leavitt, and E. Freire. The linkage between protein folding and functional cooperativity: Two sides of the same coin? *Annual Review of Biophysics and Biomolecular Structure*, 31(1):235–256, 2002.

[187] R. J. Hawkins and T. C. B. McLeish. Coupling of global and local vibrational modes in dynamic allostery of proteins. *Biophys. J.*, 91(6):2055–2062, 2006.

[188] M. Jurica, A. Mesecar, P. J. Heath, W. Shi, T. Nowak, and B. L. Stoddar. The allosteric regulation of pyruvate kinase by fructose-1,6-bisphosphate. *Structure*, 6:195–210, 1997.

[189] A. Mattevi, G. Valentini, M. Rizzi, M. L. Speranza, M. Bolognesi, and A. Coda. Crystal structure of escherichia coli pyruvate kinase type i: molecular basis of the allosteric transition. *Structure*, 3(7):729–741, 1995.

[190] T. Noguchi, H. Inoue, and T. Tanaka. The m_1 and m_2-type isoenzyme of rat pyruvate kinase are produced from the same gene by alternate rna splicing. *J. Biol. Chem.*, 261(29):13807–13812, 1986.

[191] Picture courtesy of L. Tulloch.

[192] I. Ernest, M. Callens, F. R. Opperdoes, and P. A. M. Michels. Pyruvate-kinase of leishmania-mexicana mexicana cloning and analysis of the gene, overexpression in escherichia-coli and characterization of the enzyme. *Molecular and Biochemical Parasitology*, 64:43–54, 1994.

[193] R. H. E. Friesen, R. J. Castellani, J. C. Lee, and W. Braun. Allostery in rabbit pyruvate kinase: development of a strategy to elucidate the mechanism. *Biochemistry*, 37:15266–15276, 1998.

[194] A. Fenton and J. B. Blair. Kinetic and allosteric consequences of mutations in the subunit and domain interfaces and the allosteric site of yeast pyruvate kinase. *Arch.Biochem.Biophys.*, 397(1):28–39, 2002.

[195] T. G. Consler, S. H. Woodward, and J. C. Lee. Effects of primary sequence differences on the global structure and function of an enzyme: a study of pyruvate kinase isozymes. *Biochemistry*, 28:8756–8764, 1989.

[196] A. Mattevi, M. Bolognesi, and G. Valentini. The allosteric regulation of pyruvate kinase. *FEBS Journal*, 389:15–19, 1996.

[197] D. J. Rigden, S. E. V. Phillips, P. A. M. Michels, and L. A. Fothergill-Gilmore. The structure of pyruvate kinase from leishmania mexicana reveals details of the allosteric transition and unusual effector specificity. *Journal of Molecular Biology*, 291(3):615–635, 1999.

[198] J. Wooll, R. Friesen, M. White, S. Watowich, R. Fox, J. Lee, and E. Czerwinski. Structural and functional linkages between subunit interfaces in mammalian pyruvate kinase. *J.Mol.Biol.*, 312:525–544, 2001.

[199] G. Valentini, L. Chiarelli, R. Fortin, M. Speranza, A. Galizzi, and A. Mattevi. The allosteric regulation of pyruvate kinase. *J. Biol. Chem.*, 275:18145–18152, 2000.

[200] G. Vriend. WHAT IF: a molecular modelling and drug design program. *J. Mol. Graph.*, 8:52–56, 1990.

[201] L. B. Tulloch, H. P. Morgan, V. Hannaert, P. A. M. Michels, L. A. Fothergill-Gilmore, and M. D. Walkinshaw. Sulphate removal induces a major conformational change in leishmania mexicana pyruvate kinase in the crystalline state. *J. Mol. Biol.*, 383:615–626, 2008.