

Jagiellonian University
Faculty of Physics, Astronomy and Applied Computer Science

Computational modelling of neck linker's polymeric properties
with comparison to mutated protein sequences in the context of
kinesin's overall dynamical properties

Michał Świątek
PhD Thesis

Supervisor: dr hab. Paweł F. Góra



June 14, 2024

Podziękowania

Chciałbym serdecznie podziękować prof. dr hab. Ewie Gudowskiej-Nowak i dr hab. Pawłowi F. Górze za ich nieocenioną pomoc, życzliwość i nieograniczoną cierpliwość okazane podczas powstawania tej pracy.

Abstract

This doctoral thesis is chiefly about biopolymers and rapid perturbations they are subjected to, in the context of greater biological systems of which they are functional elements. Neck linker, a biopolymer that served as a basis of my inquiries, is a critical part of kinesin, a motor protein that drives anterograde axonal transport. Neck linker is but a small part of kinesin sequence, its length does not reach 20 amino acid residues, and yet it is a part upon which the function of kinesin hinges. It is known that neck linker's contribution to kinesin's walking along microtubule involves switching conformational states. In fact, that process shares many similarities with pulling experiments, in which biopolymers are stretched while the system is far from equilibrium. Thanks to Jarzynski equality, it is possible to derive free energy differences for these processes.

This thesis presents a complex approach to biopolymers. Normal Mode Analysis is used to assess neck linker's dynamical properties in context of the native kinesin structure. Molecular Dynamics simulations are employed to carry out a protocol designed to approximate a large ensemble of non-equilibrium neck linker pulling experiments. Jarzynski equality serves to derive free energy difference of neck linker's switching between conformational states. Several 'mutated' peptides are derived from original neck linker and subjected to the same protocol. A connection has been found between point mutations and free energy difference of conformational state switching.

The thesis is divided into five chapters.

First chapter states the objectives of the thesis, introduces the biological context, within which kinesin motor protein exist, the basics of its function and history of its discovery. Then it goes on to touch upon Protein Data Bank as a source of native structures, neck linker's role in kinesin's stepping and how well defined is neck linker's sequence. Finally it reframes neck linker as a polymer, briefly describes basics of pulling experiments and hints at the concept of free energy as well as Jarzynski equality

Second chapter introduces the concept of Normal Mode Analysis and Domains Analysis and presents results of applying these methods to the native structure of kinesin obtained from Protein Data Bank.

Third chapter introduces the principles of Molecular Dynamics and defines several associated concepts relevant to subsequent simulations. It goes on to describe a simulation protocol designed in order to acquire an unusually large ensembles of stretching simulations and presents results of relevant trajectory analysis, focusing on comparison between neck linker and 'mutated' peptides.

Fourth chapter delves into theoretical descriptions of several statistical properties of a system, including free energy and its difference. It then introduces Jarzynski equality and highlights its implications for pulling experiments. What follows is an introduction of Crooks fluctuation theorem. After that, practical limitations of Jarzynski equality are mentioned and other estimators, *i.e.* the Bennett estimator, are also introduced. Lastly, the chapter discusses implications of the free energy difference between various conformational states for neck linker's biological function, makes comparison between the neck linker and 'mutated' peptides and assesses how the free energy profile of the neck linker is influenced by amino acid substitutions.

Fifth and final chapter summarizes the investigations done for the preparation of this thesis, highlights the most important results and hints at avenues for further research.

Streszczenie

Poniższa praca doktorska dotyczy biopolimerów i ich przemian strukturalnych w odpowiedzi na gwałtowne perturbacje, którym mogą podlegać jako funkcjonalna część większych systemów biologicznych. Kinezyrna to białko 'kroczące' po mikrotubulach i umożliwiające szybki anterogradowy transfer aksjonalny. Charakterystycznym funkcjonalnym elementem jej struktury, jest tzw. *neck linker*, 'łącznik' pomiędzy szyją kinezyrny a jej głową. Jest to krótki fragment łańcucha, którego długość nie osiąga nawet 20 aminokwasów, mimo tego, że jego obecność jest kluczowa dla funkcjonowania tego białka motorycznego. Udział tego 'łącznika' w krokach kinezyrny jest wyrażany poprzez przeskakiwanie pomiędzy odpowiednimi stanami konformacyjnymi. Zjawisko to posiada wiele podobieństw do eksperymentów rozciągania biopolimerów przy użyciu metod typu AFM, czy też szczypiec optycznych, w których układ znajdował się daleko od równowagi termodynamicznej. Dzięki równaniu Jarzyńskiego, jesteśmy w stanie wyznaczać różnicę w energii swobodnej pomiędzy stanami konformacyjnymi tego 'łącznika'.

Analiza drgań normalnych została wykorzystana do oceny własności 'łącznika' w kontekście natywnej struktury kinezyrny. Przeprowadzono symulacje Dynamiki Molekularnej w ramach protokołu, który został opracowany w celu uzyskania przybliżonego zespołu statystycznego dla eksperymentów rozciągania 'łącznika'. Równanie Jarzyńskiego zostało wykorzystane do określenia zmian energii swobodnej dla przeskakiwania między stanami konformacyjnymi. Protokół zastosowano także dla grupy 'zmutowanych' peptydów, uzyskanych poprzez modyfikację natywnej struktury 'łącznika'. Znalaziono zależność między obecnością mutacji punktowych w sekwencji 'łącznika' a różnicą energii swobodnej dla przeskakiwania pomiędzy stanami konformacyjnymi.

Praca składa się z pięciu rozdziałów.

Pierwszy rozdział zawiera omówienie zagadnień istotnych dla tematyki pracy. Składa się on z siedmiu części. W pierwszej części przedstawione zostały cele pracy. Druga zawiera historię procesu odkrywania białek motorycznych. Trzecia tłumaczy rolę kinezyrny w transporcie wewnątrzkomórkowym i wzmiankuje teoretyczne modele ruchu kinezyrny. W czwartej części przybliżone zostają parametry struktur zdeponowanych w *Protein Data Bank*. Piąta część zawiera przybliżenie sekwencji 'łącznika' w kontekście całej cząstki kinezyrny. Część szósta wprowadza interpretację biopolimerów jako mechanicznych łańcuchów. Finalnie, w części siódmej przedstawiony zostaje koncept badania zmiany energii swobodnej podczas rozciągania polimerów.

W drugim rozdziale, podzielonym na dwie części, wprowadzone zostają teoretyczne podstawy analizy drgań normalnych i analizy domenowej. Następnie przedstawione zostaje, jak można użyć tych metod do wyróżnienia oddzielnych, relatywnie sztywnych, domen w strukturze białka. Kolejna część zawiera dyskusję rezultatów analizy i pokazuje, że podział na domeny w strukturze kinezyrny przebiega zgodnie z oczekiwaniami, wynikającymi z funkcji tego białka.

W trzecim rozdziale przybliżone zostają podstawy Dynamiki Molekularnej, a także pojęcia i metody użyte w protokole zastosowanym do przeprowadzenia symulacji rozciągania 'łącznika' i jego 'mutantów'. Następnie pojawia się opis samego protokołu, umożliwiającego generowanie obszernych zespołów statystycznych dla procesu rozciągania peptydów. Finalnie, przedstawione są rezultaty analizy trajektorii, z naciskiem na różnice pomiędzy 'łącznikiem' a 'zmutowanymi' peptydami.

Rozdział czwarty zawiera teoretyczny opis najważniejszych statystycznych własności termodynamicznego układu, wliczając w to różnicę energii swobodnej. Rozdział zawiera też omówienie równania Jarzyńskiego i twierdzenia fluktuacyjnego Crooksa oraz wynikających z nich konsekwencji dla rozciągania biopolimerów. Uwaga zwrócona zostaje na ograniczenia w wyznaczaniu różnicy swobodnej energii za pomocą równania Jarzyńskiego i zdefiniowane zostają alternatywne estymatory różnicy swobodnej energii np. estymator Bennetta. Rozdział jest zwieńczony dyskusją na temat znaczenia różnicy energii swobodnej między stanami konformacyjnymi dla biologicznej funkcji 'łącznika'. Dokonane zostało także porównanie między 'łącznikiem' a 'zmutowanymi' peptydami i ocena, jak substytucja danego typu aminokwasów w sekwencji wpływa na profil energetyczny 'łącznika'.

W piątym i ostatnim rozdziale następuje podsumowanie pracy badawczej, która towarzyszyła przygotowaniu tej pracy doktorskiej. Podkreślone zostają najważniejsze rezultaty i zasugerowany zostaje możliwy kierunek dalszych badań.

Declaration

Parts of the results of this thesis have been previously published in my papers, in *Acta Physica Polonica B*[\[71\]](#) as well as in *Scientific Reports*[\[96\]](#).

Contents

1	Introduction	8
1.1	The objectives of this thesis	8
1.2	First motor proteins and rise of kinesin	8
1.3	Brief description of kinesin's role in neuronal intracellular transport	9
1.3.1	Theoretical models of kinesin's movement	10
1.4	Kinesin structures in Protein Data Bank	12
1.5	Specificity of neck linker's sequence and structure	15
1.6	Neck linker as a biopolymer and experimental methods that can facilitate its stretching	15
1.7	Free energy difference between conformational states	18
2	Dynamical properties of kinesin structure	19
2.1	Methods	19
2.1.1	Normal Mode Analysis	19
2.1.2	Domain Analysis	20
2.2	Results	21
3	Neck linker stretching	25
3.1	Methods	25
3.1.1	Molecular Dynamics	25
3.1.2	Force Field Model	26
3.1.3	Algorithms for energy minimization	27
3.1.4	Implicit solvation	28
3.1.5	Steered Molecular Dynamics	28
3.1.6	Tools of Trajectory Analysis	28
3.2	Simulation Protocol	30
3.2.1	Neck linker	30
3.2.2	Poly-alanine	31
3.2.3	Point Mutations	31
3.3	Results	33
3.3.1	Equilibration and selection of starting structures for stretching simulations	33
3.3.2	Analysis of trajectories averaged over ensembles of stretched peptides	36
4	Free energy difference estimation on the basis of Jarzynski equality and Crooks fluctuation theorem	51
4.1	Methods	51
4.1.1	Defining free energy for ensembles of conformational states	51

4.1.2	Free energy difference	52
4.1.3	Jarzynski equality	53
4.1.4	Crooks fluctuation theorem	53
4.1.5	Practical considerations and alternative free energy difference estimators	54
4.2	Results	55
5	Conclusions	60

Chapter 1

Introduction

1.1 The objectives of this thesis

The aim of this thesis was to simulate a pulling experiment, which was meant to approximate neck linker's switching between conformational states during kinesin stepping, perform analogous simulated experiments for several 'mutated' peptides, and then investigate what impact the substitutions in the neck linker's structure may have on the dynamical properties of the neck linker. In order to check which, if any, quantity can be tied to the aforementioned amino acid substitutions, analyses of trajectories and free energy difference estimation have been planned.

Intermediary goal involved asserting through the use of Normal Mode Analysis in conjunction with Domain Analysis, whether the native structure of kinesin obtained from Protein Data Bank preserves its functionality and if the results have implications for functionality of neck linker extracted from that structure.

1.2 First motor proteins and rise of kinesin

In year 1985 a paper was published in *Cell* by Ronald D. Vale, Thomas S. Reese and Michael P. Sheetz that identified a new type of protein motor and named it kinesin[102]. Up to this point, already two other major types of motor proteins have been identified - dynein and myosin. Dynein was initially found in early 1960s, in *Tetrahymena pyriformis*' cilia [30, 91, 58], and soon after was proposed as a key factor in cilia/flagella bending [31, 91, 58]. Myosin was discovered even as far back as 1864 by Wilhelm Kühne, the man behind the coinage of the term *enzyme*, who recognized its association with skeletal muscle tension[37]. While all three motor proteins were classified as ATPases right from the start, kinesin's much heralded role as a carrier along microtubule, its importance for axonal transport and a probable transport direction were already inferred from the initial experiments[102]. In contrast, a presence of a dynein-like protein in axonal transport and a probable transfer direction were only asserted few years after kinesin's discovery [81]. In fact, the knowledge of kinesin itself, of its function and properties, was used in the process of making this assertion[81]. Soon after carrying capacity and directionality of both kinesin and dynein has been corroborated *in vivo*[43, 42]. As new variants of each of the original three motor proteins have continued to be discovered[19], they came to be considered as three separate protein superfamilies[40, 50, 67, 41, 37]. Considering these facts it can be argued that kinesin superfamily members have been the main focus and a reference point of the research into various facets of microtubule-based transport.

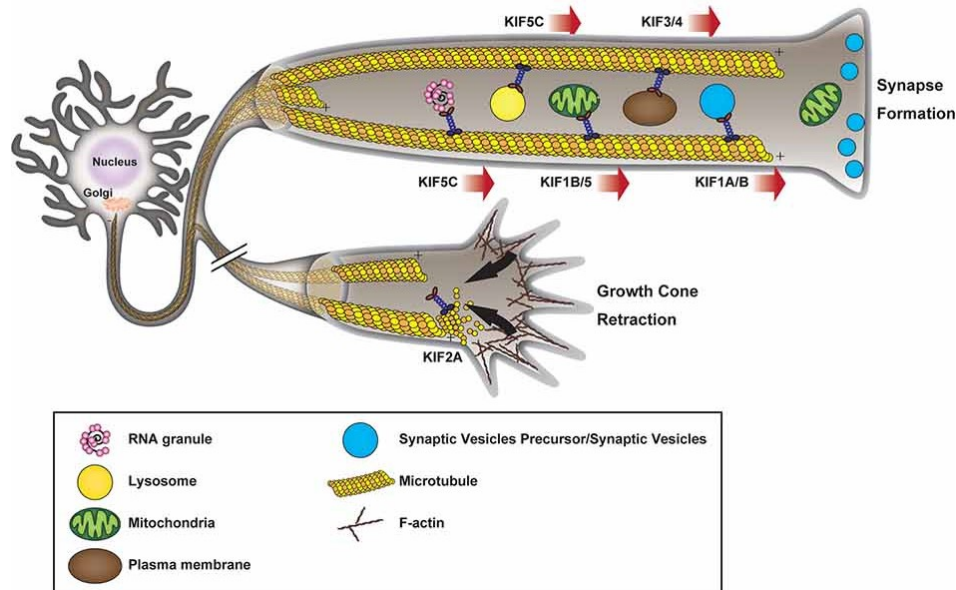


Figure 1.3.1 Microtubule-based transport contextualized within a larger scheme of neuronal intracellular transport. This figure has originally been published in a Nature review paper from 2020 by Guillaud et al[33].

1.3 Brief description of kinesin's role in neuronal intracellular transport

Kinesin's impact on cell activities is far-reaching. Using a neuron (alternatively called a nerve cell) as a basis for a description of kinesin's role in intracellular transport seems to be a very attractive proposition. It is widely known that a neuron propagates electric signals by means of cell membrane depolarization, which causes action potentials to travel along the neuron's axon. A type of junction that transfers these signals between neurons is called a synapse. Within the confines of a single signal transfer, a neuron that delivers the signals to a synapse has been assigned a descriptor *presynaptic*, while a neuron that recovers the signals from a synapse has been assigned a descriptor *postsynaptic*. In most cases a single neuron takes on a role of a postsynaptic one when receiving the signals through its dendrites and then takes on a role of a postsynaptic one when transferring the signal forward through its terminal buttons (axon culminations). The presynaptic role involves amassing vesicles filled with neurotransmitters (molecules whose release and subsequent reception on the other side of a synapse mediate a signal transfer) in the area of the terminal buttons. Anterograde axonal transport is a process by which these vesicles are gathered in one place (*anterograde* meaning *from the point of origin forward*; the point of origin in the context of microtubule being (-)end with (+)end considered the endpoint; and even more specifically in the context of a nerve cell the point of origin being near its main body and the endpoint placed closer to its terminal buttons). As had been outlined previously, kinesin (or more generally kinesin superfamily members) is a major participant of anterograde axonal transport. In fact, the experiments, which led to kinesin's identification, were conducted on an axoplasmic supernatant gathered from the squid giant axon[102].

Fig.1.3.1 (originally published in [33]¹) contains a conceptual representation of processes described above.

¹published under the Creative Commons Attribution licence (CC-BY)

We see members of kinesin superfamily (their names start with prefix **KIF**, possibly from **K**inesin **I**ncluding **F**amily; even though it is also claimed to simply mean Kinesin Superfamily[67]) dragging various types of cargo along microtubules. The general outline of kinesin structure is shown as having globular domains on each side of a long helix knot. Microtubules' nature as hollow polymeric rods of fluctuating length is also hinted at. It should come as no surprise that they are among main components of cytoskeleton. Their significance as transport tracks in eukaryotic cells has been made evident, but it is worth noting that they majorly contribute to maintaining cell shape as well as to altering it when required. In fact all these functions are very much linked - when microtubules stretch cell boundaries towards certain direction, they simultaneously make themselves available as avenues for delivery of molecules required for cell processes in newly covered space. Dimers of tubulin molecules (one α -tubulin and one β -tubulin for each pair) are working as microtubules' building blocks. They attach themselves to a microtubule in an organized manner - β -tubulin monomer is always on the outside at the (+)end of a microtubule. Kinesin bonds with β -tubulin[94] and walks in a hand-over-hand manner towards the (+)end of a microtubule[105]. Periodicity of a microtubule enforced by tubulin dimers influences the mechanics of kinesin movement. Each step is of the same size, i.e. about 8 nm for the movement of the center-of-mass. It has been previously mentioned that kinesin belongs to ATPases (enzymes functioning as catalysts for ATP - Adenosine TriPhosphate decomposition through hydrolysis). Energy gain from successful hydrolysis of a single ATP molecule is transformed by kinesin into one step towards the (+)end of the microtubule and only extremely rarely towards its (-)end. The lack of symmetry between the frequencies of forward and backward steps has long been considered to be tied to the characteristics of kinesin structure. A *structure* of a molecule refers to a relative placement of molecule's atoms in three dimensional space. This configuration is also often called a *conformation* or a *conformational state* (when focusing on impermanence of these configurations). For proteins, it is generally assumed that these configurations occur in nature. Initial insights into a kinesin structure have appeared in 1987[6] and the additional details have continued to be uncovered over the years[103, 88, 53, 63]. The kinesin found in neurons, the first one to be discovered[102] and often called *conventional kinesin*, belongs to kinesin-1 family[67]. In Fig.1.3.2 (originally published in [71]²) we have a conceptual representation of a single molecule of this type. The representation focuses on presenting differences in structure and function within kinesin chains. It highlights a presence of a two motor domains (usually referred to as *heads*), their mutually exclusive connection to the microtubule[63], a stalk providing a bridge to a cargo and a pair of elastic coils (usually referred to as *neck linkers*), connecting the heads with the stalk.

1.3.1 Theoretical models of kinesin's movement

Since motor proteins draw comparisons to mechanical motors, it is very tempting to try and abstract its biological complexity into a more manageable model. One alley for such ideas leads to redefining different parts of a motor as modules that can be differentiated by either their function, conformation or both. It would then be assumed that these modules move at different times and speeds [45]. Such considerations led to the situation where molecular motors provide a desirable testing ground for application of various theoretical models, that aim to filter the inherent complexity of biological systems [61, 62, 97, 49]. Kinesin, as a prime member of the group, is of course at the center of these movements *in vitro* [61] and *in silico* [61, 97, 107, 49].

Many theoretical models attempted to tackle the problem of kinesin walk along microtubule from various perspectives.

Two categories of the theoretical models attempt to explain behaviour of kinesin motor holistically, treating a kinesin molecule as a unit to which a set of reactions is assigned. Triggers for these reactions are also defined. These are mechanochemical models and ratchet models.

The mechanochemical models use the concept of the Markov chain, kinetic description of the reaction cycles that cause the conformational transitions, which in turn lead to mechanical movement. The kinesin motor is

²published under the Creative Commons Attribution licence (CC-BY 4.0)

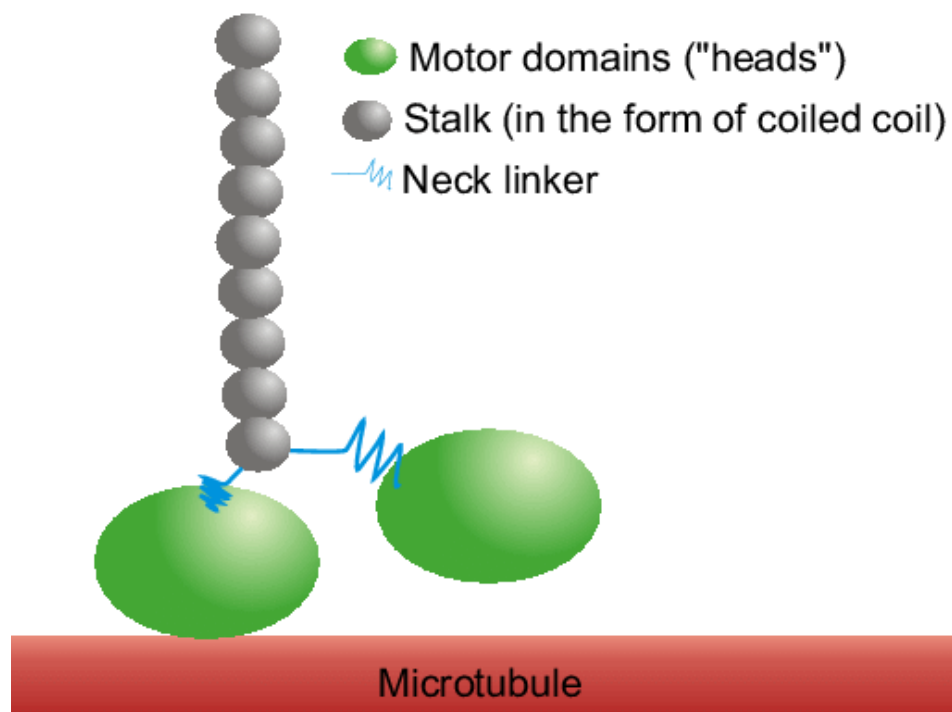


Figure 1.3.2 Conceptual representation of a single kinesin molecule focusing on differentiating structure elements based on their mechanical function. This figure has originally been published in our work[71].

treated as catalysts in these reactions. The mechanochemical models derive their parameters from the results of experiments, in which the enzymes' reaction cycles can be sufficiently analyzed. If one obtains knowledge of all the reactions that may occur on the active sites of kinesin's motor domains, one can attempt to estimate the rate of conformational transitions. One can then reach quantitative description of the mechanochemical coupling - dependence between the energy produced by the reactions and the conformational changes and how these lead to directed walking along microtubule.

The ratchet models attempt to explain kinesin's walking movement through a collection of Langevin equations and use a pair of conjoined Brownian particles moving in a periodic potential as an abstract, functional representation of a kinesin molecule. These models consider the influence kinesin's size and overcrowded environments in living eukaryotic cell have on their dynamics.

Another category of theoretical models constitute the field of **Molecular Dynamics (MD)**. In contrast to the previously summarised models, **MD** try to simulate changes in a self contained ensemble as time flows, where all relevant molecular entities are represented in some way and thermodynamic properties are kept at levels reasonably close to physiological ones. In order to prepare the ensemble for simulation it is usually required to at least acquire a chemical formula (in case of proteins, a protein sequence) of a molecule but an experimentally derived structure is even more desirable. Indeed, this is the approach that was adopted while performing research for this thesis. There are guiding examples available of the description of motor's response to ATP binding and hydrolysis acquired by **MD** methods[92, 48, 18, 48].

1.4 Kinesin structures in Protein Data Bank

One searching for a kinesin structure to study will inevitably arrive at **Protein Data Bank** (often shortened to **PDB**, however the abbreviation is also used when referring to specific structures stored in the Bank or as an extension in names of computer text files containing these structures). It can be accessed under following internet addresses: <http://www.wwpdb.org> houses the global unified version; <https://www.rcsb.org> directs to the original American one. Since its inception in 1970s[23] and later becoming universally available online[9, 10] **PDB** has become a primary, some may be even tempted to say *only*, source of protein structure data. Data stored in **PDB** are generally acquired through various experimental methods, among which big three are: x-ray crystallography, nuclear magnetic resonance spectroscopy (*NMR*) and electron microscopy. Protein structures solved through electron microscopy used to be at a distant third place[15] but are recently rapidly growing in number[13], competing with *NMR* for second spot and seem likely to burst ahead, considering the rate with which they are being reported[14]. A prevalent method, responsible for obtaining overwhelming majority of the structures is X-ray crystallography[15, 13, 14], which was first used for that purpose over a decade before the start of **PDB**[57]. Incidentally, the structure solved at that time belonged to myosin, another previously mentioned motor protein. Structures describing parts of kinesin started to appear in **PDB** over a decade after the initial discovery[53, 89, 63].

Two experimental parameters of X-ray crystallography due some of our attention are temperature and resolution. In most cases temperature of X-ray crystallography is kept at cryogenic levels (overwhelmingly in range of 90 – 100K, when it comes to kinesin structures, see Fig.1.4.2) rather than at room temperature. Advantages can be very mundane ones - it is far easier to store and move containers with crystals inside at low temperatures. As a result data can be collected remotely[26]. The advantages can also be more profound - cryogenic temperatures slow down diffusion of free radicals that may appear in a sample due to radiation[26]. Some flaws in regard to this trend are however readily apparent. One would intuitively expect that behaviour of a folded protein chain would be different in low temperatures as opposed to temperatures in which physiological processes occur. Indeed, such doubts has been voiced and a call to try and increase the number of structures solved in room temperature is being made[28, 26]. Nevertheless, while some temperature-dependent changes

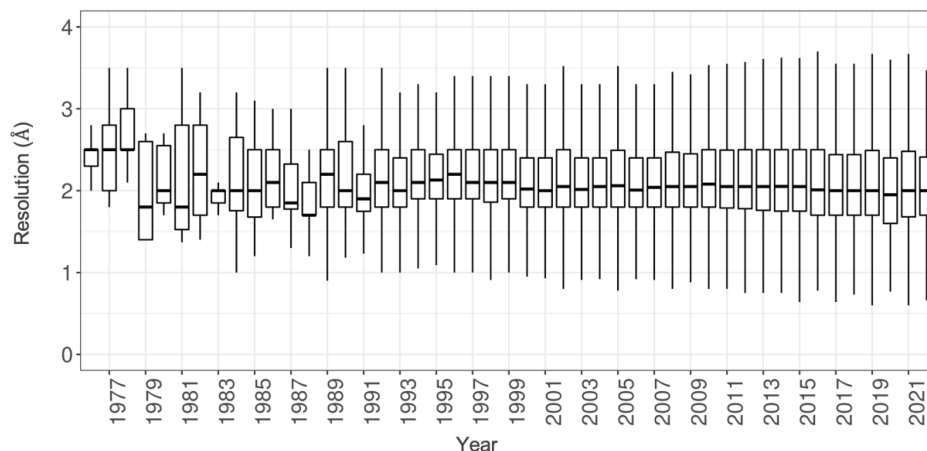


Figure 1.4.1 A set of boxplots visualizing how resolution of X-ray crystallography data in **PDB** is distributed. We can see that the distribution is more or less stable throughout the years, regardless of improvements in the field of X-ray crystallography. This figure has originally been published in [14].

in conformational states has been shown[28], as of now there are relatively few room temperature structures.

Another thing that needs to be addressed is a possible confusion when it comes to temperature in relation to **PDB** data. In **PDB** files each atom has a parameter assigned to it called *temperature value* (or *B-factor*, *temperature factor* etc.). Despite the name, this parameter is not tied to the actual temperature of the experiment but rather describes local disorder[7]:

$$B_i = \frac{8}{3} \pi^2 \langle (\Delta R_i)^2 \rangle . \quad (1.1)$$

B_i refers to B-factor for an i – *th* atom and $\langle (\Delta R_i)^2 \rangle$ refers to its mean square displacement. It allows us to make initial judgements, which regions are more conformationally active in comparison to the rest of a protein sequence.

Resolution in X-ray experiments is the shortest distance between two observed features at which they still can be differentiated. It can be considered a measure of average disorder. Resolution of separate structures in **PDB** vary, but median for the whole database is around 2.0 Å [14] (see Fig.1.4.1, originally published in [14]³). On the other hand, median for kinesin data only hovers around 2.6 Å. We can see the resolution data spread in Fig.1.4.2. The difference seems to confirm that an average kinesin molecule is more fluid than an average protein crystal structure. That is well within expectations, after all not many proteins regularly perform rapid shifts in conformational state in response to an impulse. Kinesin motors are compelled to respond with a step when an ATP molecule initiates interaction.

Final remarks regarding **PDB** structures concern missing elements in protein chains. It is a very common occurrence that a number of either whole residues or individual atoms have not been located during the experiments. It may mean that the protein chains will be truncated or even in pieces. When perusing the structures in broadly defined molecular modelling (e.g. **MD**), it is a good practice to be aware of missing elements in case they are shown to influence modelling results.

³published under the Creative Commons Attribution licence (CC-BY 4.0)

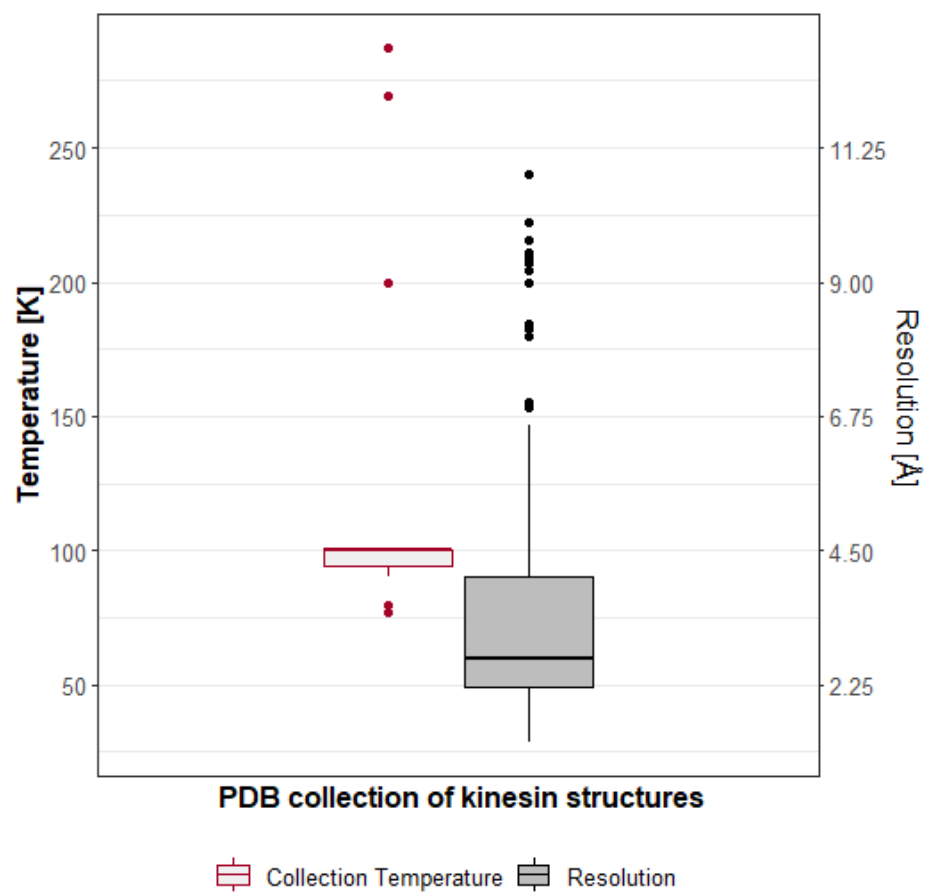


Figure 1.4.2 A distribution of X-ray crystallography data collection's temperature on the left and a distribution of its resolution on the right, plotted for the whole dataset of kinesin structures in **PDB**

1.5 Specificity of neck linker's sequence and structure

The approaches towards explaining kinesin mechanics, that are more granular than either Ratchet or mechanochemical models, inevitably arrive at the issue of neck linker. It has been shown that neck linker's change in conformation is coordinated with a step along microtubule and ATP hydrolysis[87]. More precisely, it has been shown that the *neck linker* switches from unstructured, flexible conformational state to a structured and docked conformation state after **ATP** binding. It provides crucial conformational change in the kinesin, associated with the subsequent stepping[99].

In a kinesin molecule of a prevalent type two twin amino acid chains form the motor domain in a form of twin kinesin heads. N-terminus of each chain is buried deep at the core of each kinesin head. Each head aggregates around a single Mg^{2+} ion. Each kinesin head spans 300 residues or more. Towards the C-terminus, kinesin head gives way to that part of the sequence, which forms the stalk and the chain is stabilised in a helical configuration. The helices from twin chains coil around one another.

For the purposes of this thesis, it will be useful to dive deeper into specifics concerning neck linker's chemical representation. The length of neck linker and concurrently its placement in a protein sequence varies (its C-termini and N-termini cannot be assigned an immutable index within a sequence)[82], often depending on the family to which a given kinesin molecule belongs[35]. It has been shown by Hariharan and Hancock [35] (see Fig.1.5.1, that was originally published therein) that matching patterns are present in kinesin sequence and that they correspond to the region responsible for linking a head with a stalk. High similarity across the protein family, that is plainly visible in Fig.1.5.1 (originally published in [35]⁴), hints at the possibility that the neck linker sequence is promoted through evolutionary mechanisms. If so, it is far more likely that neck linker sequences are specific combinations of amino acids that provide concrete functional benefits to the whole, rather than a random selection that endures only by accident. The relationship between the prevalent symbols for amino acids and the alternative symbols showing up in Fig.1.5.1 is presented in Table 1.1. The prevalent symbols will be in use throughout this thesis.

All in all, neck linker is a sequence within protein's amino acid chain that includes no more than 20 residues (a *residue* is an alternative term for an amino acid; it brings attention to the fact that amino acid is a part of a larger protein chain). Neck linker is sometimes considered a part of kinesin head, sometimes a neighbouring independent agent. It connects α -6 helix in coiled-coil dimerization domain and α -7 helix in the core motor domain[35]. On average, neck linker typically falls in between the 338th and the 354th kinesin residue [35] and has a mass of about 2kDa (an atomic mass unit of 1 dalton (*Da*) corresponds to $1.661 \times 10^{-27}kg$).

Ever growing bulk of data collected from experiments *in vitro* already made it clear that the rapid transition of the *neck linker* from a disordered (random coil) state to an ordered (β -sheet) conformation is a key factor in determining a mechanism of force-generation. Neck linker is thus not just a useful yet passive bystander when kinesin performs a step, but is a crucial element in dynamic processes that drive kinesin's affinity for one directional movement[35, 87].

1.6 Neck linker as a biopolymer and experimental methods that can facilitate its stretching

A term *biopolymer* is being used here to refer to molecules present in living organisms that behave like polymers or rods. If protein chains are short enough, they can be interpreted as slender rods. Such a rod should follow a beam equation:

$$Mb = fr \frac{1}{C_v}, \quad (1.2)$$

⁴published under the Creative Commons Attribution licence (CC-BY 4.0)

Kinesin-1 Family																
	311	310	320	330	340	350	360	370	3							
HsKHC	297	RTTF	ICCS	SSVND	AETKSTL	MFQ	QRAKTI	IKNTAS	VNLELTAE	QWKKYK	EKEK	TKA	QETI	AK	LE	AE
RnKHC	284	RTTIV	ICCS	SPSVN	DAETKSTL	MFQ	QRAKTI	IKNTAS	VNLELTAE	QWKKYK	EKEK	TKA	QETI	AK	LE	AE
DmKHC	303	RTTIV	ICCS	SPSVN	DAETKSTL	MFQ	QRAKTI	IKNTAS	VNLELTAE	QWKKYK	EKEK	TKA	QETI	AK	LE	AE
MmKIF5a	297	RTTF	ICCS	SSVND	AETKSTL	MFQ	QRAKTI	IKNTAS	VNLELTAE	QWKKYK	EKEK	TKA	QETI	AK	LE	AE
MmKIF5c	297	RTTIV	ICCS	SPSVN	DAETKSTL	MFQ	QRAKTI	IKNTAS	VNLELTAE	QWKKYK	EKEK	TKA	QETI	AK	LE	AE
MmKIF5b	295	RTTIV	ICCS	SPSVN	DAETKSTL	MFQ	QRAKTI	IKNTAS	VNLELTAE	QWKKYK	EKEK	TKA	QETI	AK	LE	AE
SpKHC	295	RTTIV	ICCS	SPSVN	DAETKSTL	MFQ	QRAKTI	IKNTAS	VNLELTAE	QWKKYK	EKEK	TKA	QETI	AK	LE	AE
NcKHC	300	RTTIV	ICCS	SPSVN	DAETKSTL	MFQ	QRAKTI	IKNTAS	VNLELTAE	QWKKYK	EKEK	TKA	QETI	AK	LE	AE
NhKin1	299	RTTIV	ICCS	SPSVN	DAETKSTL	MFQ	QRAKTI	IKNTAS	VNLELTAE	QWKKYK	EKEK	TKA	QETI	AK	LE	AE
UmKin2	303	RTTIV	ICCS	SPSVN	DAETKSTL	MFQ	QRAKTI	IKNTAS	VNLELTAE	QWKKYK	EKEK	TKA	QETI	AK	LE	AE
Consensus	311	RTTIV	ICCS	SPSVN	DAETKSTL	MFQ	QRAKTI	IKNTAS	VNLELTAE	QWKKYK	EKEK	TKA	QETI	AK	LE	AE
										Neck Linker						
										ABCDEFGHIJABCDEFGHI						

Figure 1.5.1 This pattern matching has originally been presented by Hariharan et al in 2009[35]. The original figure has been truncated in order to focus only on kinesin families 1 and 2, since the molecules from these two families have been employed in research summarised in this thesis. The symbols denoting amino acids are explained in Table 1.1

Amino acid	Prevalent Symbol	Alternative Symbol
Alanine	<i>Ala</i>	A
Arginine	<i>Arg</i>	A
Asparagine	<i>Asn</i>	N
Aspartic acid	<i>Asp</i>	D
Glutamine	<i>Gln</i>	Q
Glutamic acid	<i>Glu</i>	E
Glycine	<i>Gly</i>	G
Isoleucine	<i>Ile</i>	I
Leucine	<i>Leu</i>	L
Lysine	<i>Lys</i>	K
Proline	<i>Pro</i>	P
Serine	<i>Ser</i>	S
Threonine	<i>Thr</i>	T
Valine	<i>Val</i>	V

Table 1.1: A look into amino acid naming conventions. Most of the amino acids included here are the building blocks of the conserved *neck linker* sequences in kinesin families 1 and 2.

where Mb is a bending moment, Cv is a radius of curvature and fr is a coefficient controlling rigidity. An important parameter for any chain is persistence length. Persistence length is a measure of rigidity in the sense that it is in opposition to thermal fluctuations. The model that is most often used to describe chains with spring like behaviour is the **F**reely **J**ointed **C**hain (**FJC**) model. It can be described as a chain of n equally long segments (length b). Its spring-like behaviour arises from the fact that its stiffness does not converge to zero for small forces F but rather:

$$F = \left(\frac{3k_B T}{nb^2} \right) \langle R \rangle \quad (1.3)$$

with k_B being Boltzmann constant, T being temperature and R being end-to-end distance. As such the **FJC** can be considered an entropic spring, since its coefficient is kept up by temperature. The **W**orm-like **C**hain (**WLC**) model [11] is a natural continuation of **FJC** model, when length of a chain becomes considerably greater than persistence length. It has received some criticism [65, 79], but it is still considered to be an extremely good fit for describing a bending process of semi-rigid chemical chains, proteins included, since their stiffness is also temperature-dependent. The key relation between that force F and the extension x of the chain can be given as:

$$F = \frac{k_B T}{L} \left[\frac{1}{4} \left(1 - \frac{x}{L_c} \right)^{-2} + \frac{x}{L} - \frac{1}{4} \right], \quad (1.4)$$

where k_B is Boltzmann constant, T stands for temperature, L_c is the contour length of the polymer and L its persistence length.

Techniques like *Atomic Force Microscopy* **AFM** and *Optical tweezers* (an alternative name for optical trapping) has been used to establish robust protocols, that make it possible to sample responses of biopolymers to a variety of interactions. That, in turn, makes it possible to assess mechanical interactions and properties that characterize a given molecule[4]. According to these protocols singular biopolymers are pinned to the surface on one end and linked to a free-moving bead on the other. Experimental tools provide disturbance by performing a work on a free-moving bead. There can be many sources of applied force. The essential part of *Optical tweezers* method involves trapping the bead with laser. *Magnetic tweezers* generate force through a pair of fixed

magnets. The **AFM** depend on cantilever's stiffness for generation of forces. From the measurements of the cantilever's restoring force (spring force) in the **AFM** instrument, the information about elasticity (effective spring constants) and rupture forces can be derived [46, 80]. Cantilever's active movement happens only vertically, being fueled by *piezo – electric* forces at nano-level. This property makes **AFM** well-suited for sampling the relation between force and extension[78]. Typical optical tweezers method had a limit of generated force at 100 pN , while magnetic tweezers could reach 200 pN at a point of focus, but could not maintain it. The **AFM** however could already probe with force at $10^4 pN$ value back in 2008[78]. Since then the methods have undergone refinement and their potency increased.

The mention of **AFM** and optical trapping is not haphazard nor hypothetical in case of neck linker. An optical trap experiment[104] showcased that linker being kept in a state of tension is critical to the walking function of kinesin. Artificially extending the linker with additional amino acids has resulted in kinesin's effectiveness plummeting. Moreover, when a kinesin molecule has been altered in a way that the *neck linker* amino acid chain has been switched with an amino acid chain of a different composition, a loss of functionality of the whole motor was reported[17].

1.7 Free energy difference between conformational states

Taking into consideration how fruitful were these endeavours, a question one could ask is if one could employ theoretical methods of computational modelling to reproduce this types of experiments for kinesin's neck linker in the form of **MD** simulations and use theoretical chain models to analyze its behaviour. That question seems to have been answered affirmatively[35].

One can go a step further and wonder whether one can gain insight into inner workings of thermodynamic processes that occur, maybe even track them to a certain extent. The stretching of biopolymers by means of either **AFM** or optical tweezers leads to unfolding of their structures which can be understood as changing between conformational states. A free energy difference between these states can be considered a telling characteristic of a biopolymer in question. It has become possible to obtain a free energy difference between conformational states from stretching experiments[70, 24, 36, 20], thanks to the seminal works of Jarzynski[52] and Crooks[21]. Jarzynski formulated a relation connecting free energy difference between two states with an ensemble of non-equilibrium work measurements, each related to a singular realization of state-transition. Crooks on the other hand connected these properties in the form of the fluctuation theorem. In this thesis I have endeavoured to make use of Jarzynski and Crooks relations to find free energy differences between conformational states of neck linker.

Another step further would involve investigating the changes in these inner workings in response to changes in the kinesin sequence. That would allow us to ascertain how the presence of certain amino acids influences neck linker's dynamics. It is these dependencies that this thesis attempts to establish.

Chapter 2

Dynamical properties of kinesin structure

Before time consuming **MD** simulations, it is prudent to use a quicker method that can estimate whether chosen PDB structures do showcase neck linker's expected properties under **MD** methodology. **Normal Modes Analysis (NMA)** was employed to achieve that.

2.1 Methods

2.1.1 Normal Mode Analysis

NMA takes advantage of minimisation principles to obtain a full picture of vibrational modes of a system. It is understood that a structure that has reached energy minimum can be considered to be at temperature of 0K, which means that any kind of movement has ceased[68]. Internal energy U at temperature T can be given as

$$U(T) = U_{trans}(T) + U_{rot}(T) + U_{vib}(T) + U_{vib}(0), \quad (2.1)$$

where U_{vib} , U_{trans} and U_{rot} are vibrational, translational and rotational terms respectively. It is evident that at temperature 0 K the term $U(0)$ still contributes. Vibrational term U_{vib} can be calculated from following:

$$U_{vib}(T) = \sum_i^{N_{nm}} \left(\frac{h\nu_i}{2} + \frac{h\nu_i}{\exp\left(\frac{h\nu_i}{k_B T}\right) + 1} \right), \quad (2.2)$$

where N_{nm} is a number of normal modes, ν_i is an i^{th} vibrational frequency of a system. These frequencies can be found for a molecule with N atoms by calculating second derivatives of potential energy, creating a $3N \times 3N$ *Hessian* matrix H which is then mass weighted

$$H^* = M^{-1/2} H M^{-1/2}, \quad (2.3)$$

giving a matrix H^* . M is a $3N \times 3N$ matrix of masses of N atoms. The eigenvalues of H^* correspond to vibrational frequencies, while its eigenvectors to normal modes. In practice computation of these is not necessarily efficient, as the CPU time grows proportionally to N^3 and memory usage proportionally to N^2 [38]. If one is interested only in low-frequency modes, one can attempt some approximations[38]. It is possible to obtain matrix H^* for non-Cartesian coordinates q which span from $j = 1$ to $3N$, where each q_j can be interpreted as a degree of freedom. In order to do so, a matrix Tr must transform Cartesian coordinates

$$H_q^* = Tr H^* Tr^T \quad (2.4)$$

with H_q^* representing a mass-weighted hessian with new coordinates. The eigenvalues and eigenvectors of H_q^* still correspond to vibrational frequencies and normal modes and now we can ignore an unwanted degree of freedom by simply removing a matching coordinate q_i . That way we can control the computational load while still obtaining accurate frequency values for low-frequency modes.

2.1.2 Domain Analysis

Domain Analysis[38] starts where **NMA** ends. It utilises a reference structure and selected normal modes to estimate scale of vibrational deformations. It also leads to clustering of structures regions that exhibit similar large-scale motions. For Domain Analysis to give meaningful results, some conditions must be met. The environment of the analyzed structure must cause it to become overdamped as low-frequency modes experience anharmonic perturbations. As a result, force field's influence on large scale motion wanes.

A deformation energy definition employed by *DomainFinder* involves the interaction energy between two particles i and j of the elastic network and reads:

$$E_i = \frac{1}{2} \sum_{j=1}^N k(R_{ij}^{(0)}) \frac{|(d_i - d_j) R_{ij}^{(0)}|^2}{|R_{ij}^{(0)}|^2} \quad (2.5)$$

where i, j denote C_α atoms, $d_{i,j}$ denotes corresponding infinitesimal displacements from original positions (displacement of the atom in the mode to be analyzed), R_{ij}^0 is meant as distances between pairs of atoms i, j in a submitted structure and $k(R_{ij}^0)$ is an effective harmonic force constant, that reduces with a spatial distance. It is given by the relation

$$k(R_{ij}^{(0)}) = C \exp \left(-\frac{|R_{ij}^{(0)}|^2}{r_0^2} \right). \quad (2.6)$$

The point of this is to maintain a force field of short range (r_0), that inhibits interactions between potential domains. Parameter C has been chosen arbitrarily, as $47,400 \text{ kJ} \times \text{mol}^{-1} \text{ nm}^{-2}$ at temperature 300K, to enforce compatibility with the *Amber 94* force field. Amplitudes of $d_{i,j}$ are characterized by the following equation

$$\sum_{i=1}^N |d_i|^2 = fN, \quad (2.7)$$

with f acting as a scaling factor of value 1 nm^2 . Rotation (Ω) and translation (T) vectors have been tied to displacements d_i :

$$d_i = T + \Omega \times R_i. \quad (2.8)$$

In the equation 2.8, R_i stands for the i atom position. When displacement vectors do not describe pure rigid-body motion, linear least-squares fit is employed to determine values of T and Ω . The structure is then further divided into cubic compartments (cubic's side length is 1.2 nm). They are ignored unless they are found containing at least 3 atoms. Additionally, their average deformation energy's value should be below a pre-defined threshold. The rotation vectors as well as translation vectors of those cubes must be calculated. A definition of similarity, that has been used to identify clusters of cubes of similar mobility, is given below

$$\mathcal{S}_{ij} = 3 \frac{|\Omega_i + \Omega_j|}{|\Omega_i - \Omega_j|} + \frac{|T_i + T_j|}{|T_i - T_j|}. \quad (2.9)$$

The rotation vectors have been found empirically to contribute towards greater precision when identifying domains. Thus they are given a greater weight. In the last step, a cluster has been created, while obeying the criterion

$$\mathcal{S}_{ik} > \frac{\mathcal{S}_{ij}^{max}}{c}, \quad (2.10)$$

where c is a parameter named *domain coarseness*. The domain coarseness c governs the determination of rigid regions that exhibit uniform motion. It establishes the threshold of similarity between residues' rigid body motions that must be exceeded if residues in question are to be considered a part of the same dynamical domain.

Depending on the value at which c is set, domain clustering can be either too permissive or too restrictive[86]. Its proper scaling allows us to separate them from the intermediate regions, which are responsible for the fluctuations of the regions left on the outside of the domains. The coarseness parameter must be greater than 1 (see equation 2.10). All cubes contributing to the relation 2.10 populate a single cluster.

2.2 Results

The structure of *Kinesin Heavy Chain* has been obtained from **PDB**(id:3kin)[63]). The specific positioning of heads in 3kin structure causes the axis of the rotational symmetry of 120° to be aligned with coiled coil's axis [63]. This configuration would prevent each kinesin head from attaching itself to tubulin dimers if the other one managed to do so[63]. However, the 3kin kinesin chain is not connected to a tail domain that limits free movement[56]. At the same time, the fact that a kinesin molecule rotates around stalk axis is well known[85, 72].

DomainFinder application[39] has been employed with the aim of analyzing the structure described above. Domain Finder has the ability to execute both the approximate **NMA** described in 2.1.1 and Domain Analysis described in 2.1.2. The 3kin structure is first the basis for calculation of normal modes and then a reference structure in domain analysis.

There are choices that had to be made regarding the parameters of the domain analysis, that could significant impact on the results. Inclusion of normal modes in the domain analysis must be done carefully. Increasing their number leads to a more detailed view into dynamical properties of 3kin structure. Unfortunately, as the modes are gradually added to the set, average deformation energy also gradually increases. In effect, the modes, that contribute to rigid body motions, have their influenced gradually diminished. And the rigid body motions are what we were after. It must also be considered how to relate the set of selected normal modes to a deformation energy threshold. The role of a deformation energy threshold is to filter out molecule regions of insufficient rigidity. The sensible starting values of a energy deformation threshold proposed by Hinsen[39] range from $400 \text{ kJ} \cdot \text{mol}^{-1}$ to $500 \text{ kJ} \cdot \text{mol}^{-1}$. In general, such deformation energy threshold should be selected, that the average deformation energy of selected modes does not exceed it.

The average deformation energies of the 16th mode and the 17th mode are $393.03 \text{ kJ} \cdot \text{mol}^{-1}$ and $416.98 \text{ kJ} \cdot \text{mol}^{-1}$, respectively. One is above the threshold of $400 \text{ kJ} \cdot \text{mol}^{-1}$, while the other is below the threshold. After checking the other modes, it has become apparent that the average deformation energies of preceding modes are lower than that, while the average deformation energies of subsequent modes are higher than that. As seen in Fig.2.2.1, the distribution of deformation energies in the kinesin is heavily skewed towards the low values, when 15 modes are selected for analysis. However, if we increase the number of modes selected, then the number of low deformation energies rapidly decreases. If 18 modes are selected, the lowest deformation energies from the first bracket ($0\text{-}100 \text{ kJ} \cdot \text{mol}^{-1}$) are basically equal in number to the energies from the second bracket ($100\text{-}200 \text{ kJ} \cdot \text{mol}^{-1}$) So already we are running the risk of curtailing the impact of low energy modes. There are less than 90% regions under the threshold of $400 \text{ kJ} \cdot \text{mol}^{-1}$, if the number of selected modes is 17 or more. Taking into consideration the above, I have decided to select 16 normal modes for domain and deformation analyses. The deformation energy distribution of 16 modes still keeps the dominance of low energy

modes, so it seems like a safe choice. The deformation energy threshold has been set to $400 \text{ kJ} * \text{mol}^{-1}$. As a result of this choice, we have acquired the image presented in the upper panel of Fig.2.2.2. The c parameter has been kept at default value, 10. The resulting dynamical domain decomposition can be seen in the lower panel in Fig.2.2.2.

It can be easily noticed in the upper panel of Fig.2.2.2 that deformation energies of kinesin heads are much lower than the energies of the *neck linker* and at the beginning of the *stalk* (two intertwined helices, coiled coil). That means that kinesin heads are more rigid than other regions in the structure. The deformation energy is not distributed symmetrically between two kinesin heavy chains. One of them is missing 371st and 372nd residues. That fact was mentioned in the PDB file[63]. Since the missing residues should be located at the C-terminus of the chain, it is then not surprising to observe some small discrepancies between dynamical properties of the chains. Another factor that may lead to some discrepancies between the heads is the fact that the placement of almost all corresponding helices in the two chains is slightly shifted. A good example of are the analogous parts of the sequence of the heavy chain, *Lys283 – Gly293* (amino acid symbols can be found in Table 1.1). In one head that region constitutes one helix. In the other head, the helix is split in two (*Lys283 – Ile287, Gln289 – Gly293*). Sheets within the kinesin structure are not affected to a similar degree.

These rough observations are consistent with a more precise view provided by the domain analysis (Fig.2.2.2, lower panel). We have kept the c parameter at the default value 10, because the decomposition it provides is well aligned with deformation energy distribution and there is just no good reason to change its value. If the value of c is increased up to 54, which would make c pointlessly lenient, eventually all domains are subsumed into olive domain, but such approximation is not worthwhile.

In the lower panel Fig.2.2.2 we can see decomposition of the kinesin structure into several dynamical domains. Each domain can exhibit unique movement independently from the others. The olive colored and green colored domains are equivalent to two kinesin heads. The olive colored display the highest similarity score, while the green colored domain's similarity score is the 2nd highest. The cyan domain has got the 3rd highest similarity score. It can be identified as the *stalk*. The domain with the lowest similarity score is the blue one. It represents *neck linker*. It is a hoped for and welcome result, since the achieved domain decomposition aligned itself well with functional units of kinesin - a motor protein.

Neck linker's low score is exactly what would be expected, as it is a sequence that is, in some configurations, supposed to have a flexible structure relative to other kinesin's functional structures and experience exhibit softer motions. That is the *conundrum* associated with *neck linker*. On one hand, it has to have enough flexibility to empower the kinesin head's diffusive scanning of microtubule's surface. On the other hand, in certain biological configurations, like docking with the **ATP**-binding, *neck linker* is supposed to be more rigid. Particularly, when the kinesin heads are both in connection with microtubule and *neck linker* needs a certain level of stiffness to provide opportunity for mechanical coupling. Since *neck linker* is essentially in control of kinesin stepping motion, modelling its mechanical properties and testing them with **MD** simulations can be considered as important contributions towards gaining greater insight into its inner workings.

It has been suggested that dynamical properties of kinesins are not modeled by **NMA** as precisely as the dynamical properties of myosins[108]. The **NMA** approximation that Zheng and Doniach tested was proposed by Tirion[98]. The approximation I have employed was introduced by Hinsen and is different, because a pair force constant's distance dependance is represented exponentially. Since my domain analysis described in this of kinesin produces a sensible separation of kinesin structure into domains, I have provided support towards the notion that Hinsen's approximation is a good fit for kinesin, avoiding problems of Tirion's model.

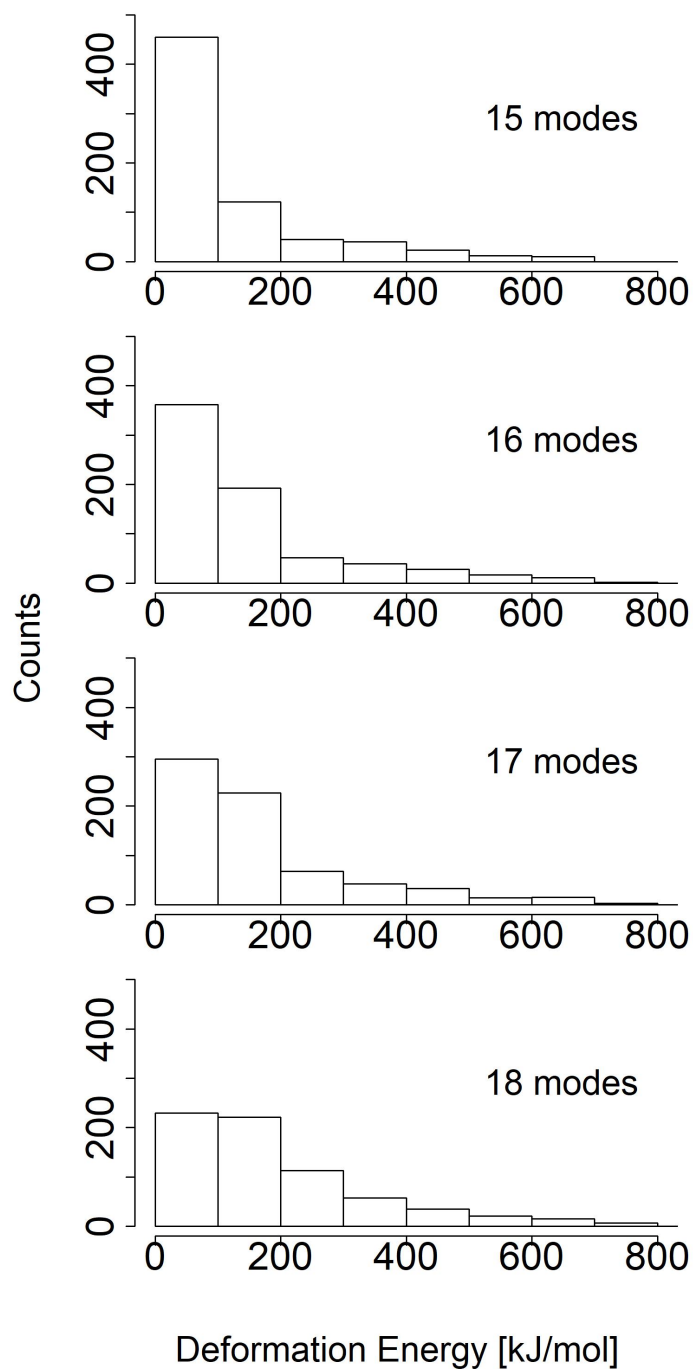


Figure 2.2.1 A comparison of deformation energy distributions in the kinesin heavy chains. The deformation energy values of each distribution have been obtained by utilising different number of low-frequency normal modes. The deformation energy threshold that differentiates between rigid and flexible region has been set at $400 \text{ kJ} * \text{mol}^{-1}$.

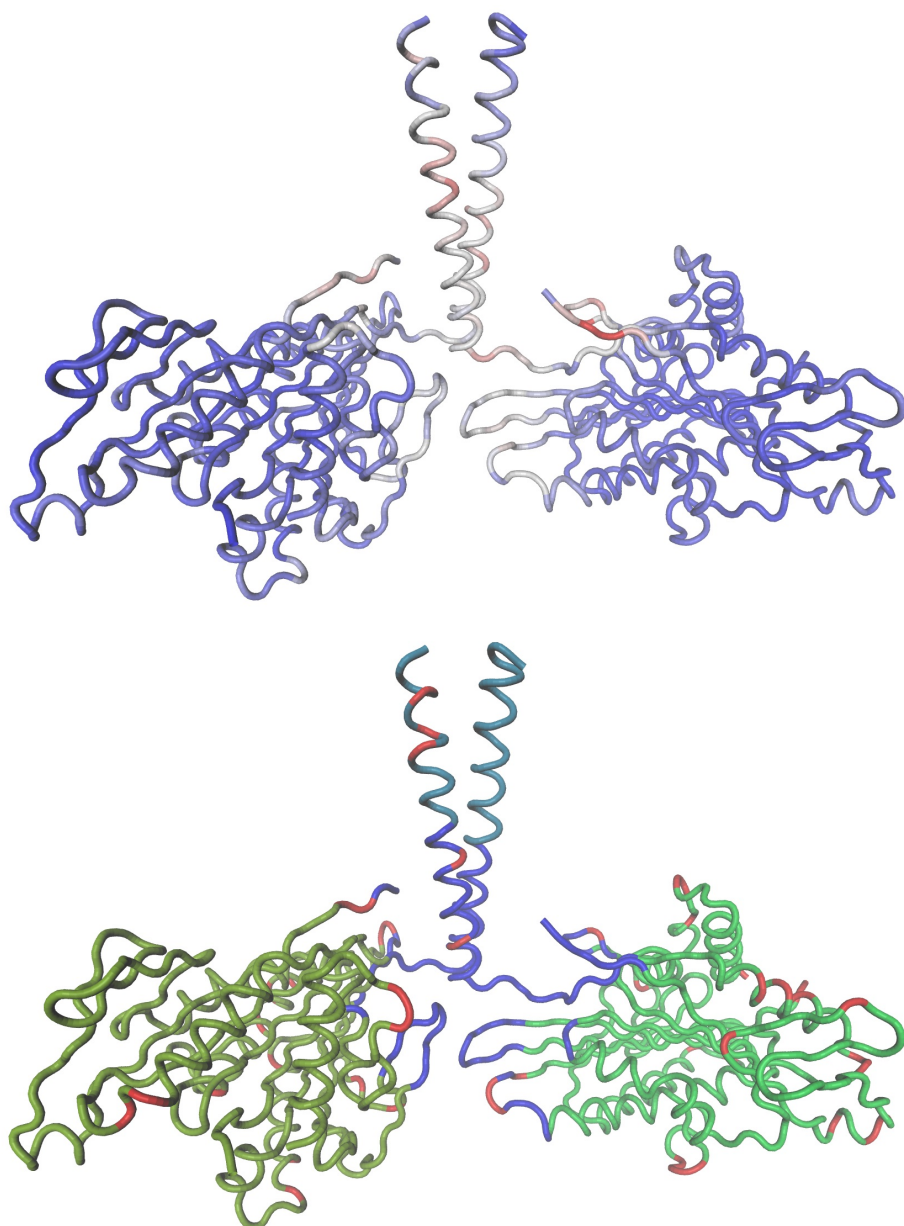


Figure 2.2.2 Distribution of deformation energies in the kinesin structure (upper panel) and derived domains (lower panel) based on 16 low-frequency modes. The deformation energy threshold has been set at $400 \text{ kJ} \cdot \text{mol}^{-1}$, domain coarseness c at 10. Upper panel: as color scale goes from blue to red, deformation energy values go from low to high. The blue color represents regions of deformation energy well below threshold, while light blue and light red parts symbolize regions slightly below and slightly above the threshold, respectively. Lower panel: each cluster of vibrational energies can be identified by its color. The domain incorporating *neck linker* region is blue, the heads are olive and green while the stalk fragment is cyan. The images has been created using VMD 1.9.3 [47] and Tachyon [95].

Chapter 3

Neck linker stretching

3.1 Methods

3.1.1 Molecular Dynamics

Molecular Dynamics (MD) take advantage of the fact that thermal motions of atoms at nanoscale are in line with classical physics. All parts of a system are given Cartesian coordinates. A force field is defined, which includes parameters of system elements and makes them more than just arbitrary points in Cartesian space. Most importantly, it includes a potential function. The purpose is to simulate trajectories of all system elements. System elements may mean atoms, or group of atoms of varying size. Molecule elements is used when referring to system elements that belong to a defined molecule. Trajectories are acquired by integrating equations representing Newton's second law of motion:

$$\frac{d^2 r_i}{dt^2} = \frac{F_{r_i}}{m_i}, \quad (3.1)$$

where m_i is a mass of a system part, r_i an i^{th} coordinate of a direction vector r and F_i is a force impacting the system part aligned with r_i . The force F_i can be obtained with the help of potential function $V(r_1, r_i, \dots, r_N)$ (N is a number of system elements):

$$F_i = -\frac{dV}{dr_i}. \quad (3.2)$$

The leap-frog integrator[44] is the default algorithm for these calculations in **GROMACS** (**G**RONingen **M**Achine for **C**hemical **S**imulations) package[8, 69]. The leap-frog algorithm takes its name from the fact that recalculations of a position vector r happen in-between recalculations of its velocity v :

$$v(t + \frac{1}{2}\Delta t) = v(t - \frac{1}{2}\Delta t) + \frac{\Delta t}{m} F(t), \quad (3.3)$$

$$r(t + \Delta t) = r(t) + \Delta t v(t + \frac{1}{2}\Delta t). \quad (3.4)$$

Even though the velocity v and the position vector r update interchangeably, the time difference Δt between these updates is constant and of the same value. The time difference Δt is often referred to as a *simulation step*. It should not be confused with a kinesin step, to which I often refer throughout this thesis. Any size of Δt can theoretically be chosen, but obviously smaller the step gets, more precise the trajectories become.

Unreasonably large steps lead to explosion of a system, but sometimes unstable systems can explode even with reasonable δt set. Molecules like proteins function in nature in relatively stable environments, as every living being maintains *homeostasis*. Additionally, a system kept at thermodynamic equilibrium allows for calculation of thermodynamic quantities. For these reasons **MD** simulations may attempt to maintain stable pressure or temperature around prescribed value among others thermodynamic quantities. **GROMACS** package offers a selection of temperature coupling and pressure coupling algorithms. It is possible to enforce schemes that lead to different kinds of thermodynamic ensembles. Canonical ensembles often referred to with acronym *NVT* - number of N elements, volume V and temperature T are tied to set unchangeable values. Microcanonical ensemble (*NVE*) arises when total energy E is kept at fixed value directly instead of controlling temperature. The isothermal-isobaric ensemble (*NPT*) in comparison to *NVT* ensemble can be achieved when control of volume v is given up in favour of keeping pressure p fixed. In order to alleviate potential problematic behaviour that could arise on the boundaries of the simulated system, *periodic boundary conditions* are usually employed in **MD** simulations. The virtual box in which system is kept is treated as if it were cloned and placed around itself, like a crystal cell within a crystal. What this means in practice is that the system elements 'touching' one side of the box interact with those on the other side. Aside from already mentioned **GROMACS**, other **MD** packages widely used for **MD** simulations are **AMBER**[16] and **CHARMM**[12].

3.1.2 Force Field Model

A potential function is needed to learn the values of the forces influencing a molecule's trajectory (see equation 3.2). The most common potential function's terms (V_{bonds} , V_{angles} , $V_{torsions}$ and $V_{non-bonded}$) depend on the force field used. Length x_i of a i^{th} covalent bond is following Hooke's law:

$$V_{bonds} = \sum_{bonds} \frac{k_i}{2} (x_i - x_{i,0})^2, \quad (3.5)$$

$x_{i,0}$ being a reference value for i^{th} bond and k_i is a force constant for i^{th} bond. An angle θ_i between neighbouring covalent bonds that share one atom follows analogous, harmonic relation:

$$V_{angles} = \sum_{angles} \frac{k_i}{2} (\theta_i - \theta_{i,0})^2, \quad (3.6)$$

where, $\theta_{i,0}$ is a reference value for i^{th} angle and k_i is a force constant for i^{th} angle. The term $V_{torsions}$ is described below:

$$V_{torsions} = \sum_{torsions} \frac{V_1^i}{2} (1 + \cos(\omega_i + f_1^i)) + \frac{V_2^i}{2} (1 + \cos(2\omega_i + f_2^i)) + \frac{V_3^i}{2} (1 + \cos(3\omega_i + f_3^i)), \quad (3.7)$$

where ω_i is an i^{th} dihedral angle, V_1^i , V_2^i and V_3^i are Fourier series' coefficients while f_1^i , f_2^i , f_3^i are phase angles. This particular shape of the term comes from **OPLS (Optimized Potentials for Liquid Simulations) all atom** force field[54]. Non-bonded term $V_{non-bonded}$ describes relations between atom pairs. Pairs of the same molecule must be separated by three covalent bonds in a chain to be counted. It contains two parts. One of these describes Coulomb potential and the other models *van der Waals* forces. The model for van der Waals forces is called *Lennard-Jones* potential:

$$V_{non-bonded} = \sum_i \sum_{j=i+1}^N \left(4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right). \quad (3.8)$$

A coefficient ϵ_{ij} describes a minimal value of Van der Waals potential between atoms with indices i and j , r_{ij} is a distance between these atoms, σ_{ij} is a distance between them when the Van der Waals potential value equals 0. The symbols q_i and q_j refer to charges of i^{th} and j^{th} atom respectively, and ϵ_0 stands for the dielectric constant. Lennard-Jones potential is the term that prevents two parts of the system, with no Coulombic interaction between them, from crashing into each other. It causes the repulsion of atoms at minimal distances, attraction at medium distances and at large distances begins to approach 0. Both Coulomb and Lennard-Jones term can then be filtered to exclude long range interactions, depending on relation between practical needs of computation and validity of the model. The filtering happens at pre-MD stage, where cut-off distance for the terms' interaction can be set.

A force field contains not only the terms, but also coefficients present in the terms and system elements' parameters. It is also at this stage that the decision is made regarding granularity of the system. In *coarse-grained* force fields atoms are united into larger system elements, which have unified parameters assigned to them, while in explicit force fields all atoms are individually parameterized (**series OPLS** all atom force field mentioned above is an example of that approach).

3.1.3 Algorithms for energy minimization

As a MD simulation starts velocities are generated according to temperature control scheme. Rapid fluctuations are applied to the system and there is a very real risk that it will fail to reach stability. Necessary steps must be taken before the start of a MD simulation. They include several rounds of energy minimization. A molecule is presented as a set of molecule elements(usually atoms), each assigned names and Cartesian coordinates. The force field then interprets the molecule. Energy minimization procedures attempt to move the system towards one of its minima, by traversing molecule's energy hypersurface. While there exist non-derivative minimization methods like e.g. *Simplex* the widely used ones, *Steepest Descents* and *Conjugate Gradient*, belong the first order class of derivative methods, in other words they peruse an energy gradient in their procedures. Steepest Descents Method follows the prescription:

$$s_i = \frac{-g_i}{|g_i|}, \quad (3.9)$$

where s_i is a unit vector of $3N$ dimensions at i^{th} iteration, which decides the direction and g_i is a gradient calculated at the beginning of i^{th} iteration from the energy function. A size of a step in the direction pointed out by s_i is pre-selected before the start of the algorithm. If a step is larger than width of an energy valley the algorithm traverses, it will start to oscillate around the bottom, without reaching it. The conjugate gradient algorithm is described below:

$$s_i = -g_i + \gamma_i \cdot v_{i-1}. \quad (3.10)$$

Here the $3N$ -dimensional direction vectors s_i and s_{i-1} are not unitary as is the case for Steepest Descents. One term of the vector s_i is of equal value to the i^{th} gradient but it is being modulated by the vector of the previous step, s_{i-1} . The weight that is assigned to it is decided by a coefficient γ_i . γ_i is formulated as:

$$\gamma_i = \frac{g_i \cdot g_i}{g_{i-1} \cdot g_{i-1}}. \quad (3.11)$$

As we can plainly see here, if the gradient of the current step is of far lesser value than the gradient of the previous step, the influence of the previous step is greatly diminished. As γ_i nears 0, the Conjugate Gradient step becomes indistinguishable from the Steepest Descents step. However if g_i has higher value than g_{i-1} , then γ_i causes the term including g_{i-1} to override directions proposed by the first term. Since the Conjugate Gradient algorithm always require a preceding step, initial steps is chosen with Steepest Descents method.

3.1.4 Implicit solvation

Usually water molecules are treated within a system explicitly as *bona fide* molecules, consisting of three system elements. In some cases it is beneficial to substitute explicit water models with implicit representation. The benefits are largely connected to computational costs, as simulations with implicit solvent use up less resources. Implicit solvent's implementation in **GROMACS** belongs to *GB/SA* class[84] - it combines *generalized Born-formalism* with *Accesible Surface Area* method. The general idea is that each dissolved system element's contribution to Gibbs free energy of solvation depends on the percentage of its surface that a solvent can reach. The calculation of the effective Born radius b of each element for Coulomb interactions is necessary:

$$G_{polar} = \left(1 - \frac{1}{\epsilon}\right) \sum_{i=1}^N \sum_{j>1}^N \frac{q_i q_j}{\sqrt{r_{ij}^2 + b_i b_j \exp \frac{-r_{ij}^2}{4b_i b_j}}}. \quad (3.12)$$

where ϵ symbolises dielectric constant, q_i and q_j are electrostatic charges on system elements, while r_{ij} is the distance between system elements i and j . Hydrophobic interactions $G_{non-polar}$ are handled by the ASA component:

$$G_{non-polar} = \sum_i^N \sigma_i S_i, \quad (3.13)$$

where σ_i is an atomic solvation coefficient that has been determined experimentally for atom type i and S_i is accessible surface area of all present atoms of type i .

A simple sum give us overall Gibbs free energy of solvation G_{solv} :

$$G_{solv} = G_{non-polar} + G_{polar}. \quad (3.14)$$

3.1.5 Steered Molecular Dynamics

Aside from energy potential function that describes all atom interactions it is possible to introduce an independent arbitrary potential. We can then select system elements that are to fall under its influence and observe how the system responds to that kind of perturbation *Steered Molecular Dynamics*. One of possible choices for potential is linear potential L between two system elements, i and j :

$$L = k r_{ij} \quad (3.15)$$

where k is a force constant and r_{ij} is a distance vector between the system elements. Obviously, a simple operation gives as constant force $\vec{F}$:

$$\vec{F} = -\nabla_r L, \quad (3.16)$$

as having equal value to coefficient k

3.1.6 Tools of Trajectory Analysis

We can define an *end-to-end* vector as a vector $\vec{R}$ between two ends of a polymer. In case of proteins, it usually means between centers of mass of the N -terminal residue and center of mass of the C -terminal residue in amino acid chain. $\vec{R}$ can be construed as:

$$\vec{R} = \sum_i^n \vec{r}_i, \quad (3.17)$$

where an i^{th} vector $\vec{r}_i$, in the context of proteins, represents one of the n vectors connecting neighbouring residues (where n is $N - 1$ if N means number of residues). If we calculate the relevant mean square end-to-end distance $\langle \vec{R}^2 \rangle$, assuming $|\vec{r}_i| = l$, we reach:

$$\langle \vec{R}^2 \rangle = \sum_i^n \vec{r}_i \cdot \sum_j^n \vec{r}_j = l^2 \sum_i^n \sum_j^n \langle \cos \theta_{ij} \rangle. \quad (3.18)$$

In case of $i = j$, the angle $\cos \theta_{ij} = 1$, so the relevant term equals $l^2 \cdot n$, which leaves us with the term

$$\sum_i^n \sum_j^n \langle \cos \theta_{ij} \rangle, i \neq j. \quad (3.19)$$

In a basic **FJC** model, where torsional movements are unrestricted and equally probable, the angle average $\langle \cos \theta_{ij} \rangle$ is 0. When torsion are restricted (e.g. **MD**), local correlations will cause the term to grow with n until reach the limit of $7 - 9n$ [90]. A conclusion is reached from these deliberations, that the end-to-end distance is dependent on length of a polymer.

Radius of gyration R_g for a molecule composed of N system elements is a measure of mass distribution around the center of mass:

$$R_g = \sqrt{\frac{\sum_i^N |\vec{r}_i|^2 \cdot m_i}{\sum_i^N m_i}}, \quad (3.20)$$

where $\vec{r}_i$ is a position vector of the i^{th} molecule element in relation to the center of mass and m_i is its mass. Radius of gyration R_g in model chains is related to mean square end-to-end distance $\langle \vec{R}^2 \rangle$ as follows:

$$\langle \vec{R}^2 \rangle = 6 \langle R_g^2 \rangle, \quad (3.21)$$

however for protein chains this does not seem to be strictly upheld[75].

Root Mean Square Deviation (RMSD) is a measure of similarity based on assessing displacement between two configurations of the same system. In the context of **MD** it compares mass-weighted position $r(t)$ at different times for each of N molecule elements:

$$RMSD(t_1, t_2) = \sqrt{\frac{\sum_i^N m_i |r_i(t_1) - r_i(t_2)|^2}{\sum_i^N m_i}}. \quad (3.22)$$

A position $r_i(t)$ of i^{th} element is given at time points t_1 and t_2 . Usually $t_1 = 0$ and it is so in our analysis. **RMSD** is not normalized in regard to the size of the molecule. A measure ρ , complementary to **RMSD**, was introduced by Mayorov and Crippen in 1995[74]. The purpose of ρ was to eliminate a size factor in distance deviation calculations. It can be formulated as:

$$\rho(r_1, r_2) = \frac{2R_g(d)}{R_g(s)}, \quad (3.23)$$

where r_1 and r_2 are two compared molecules. $R_g(d)$ and $R_g(s)$ are radii of gyration for two artificial configurations. The s_i coordinate of configuration s is defined as:

$$s_i = \frac{r_{1i} + r_{2i}}{2}, \quad (3.24)$$

while the coordinate d_i of configuration d is given as:

$$d_i = \frac{r_{1i} - r_{2i}}{2}, \quad (3.25)$$

where r_{1i} and r_{2i} are i^{th} coordinates of compared molecules 1 and 2. The typically used threshold for structures similarity with **RMSD** is around $2 - 3\text{\AA}$ ($0.2 - 0.3nm$), but with ρ it is in $0.3 - 0.5$ range (ρ itself ranges from 0 to 2 where 2 implies great dissimilarity). The last measure to be introduced is **Root Mean Square Fluctuation (RMSF)**. It is calculated for each molecule element independently and can be understood as *Standard Deviation* of its position averaged over time of simulation:

$$RMSF_i = \sqrt{\left\langle (r_i - \langle r_i \rangle)^2 \right\rangle}, \quad (3.26)$$

where r_i is a position of molecule element r at time i (i is in range $1...N$ where N is a number of time steps).

3.2 Simulation Protocol

3.2.1 Neck linker

I have perused the **PDB** Databank (id:3b6u)[9] in order to acquire a structure of a Kinesin-2 family's member, a kinesin like protein *KIF3B*[67]. A sequence of 19 amino acids, 17 of which are considered to be a functional part known as a *neck linker*, has then been extracted from the motor domain found in the **PDB** file. The structure has been optimized first with the *Steepest Descents* algorithm and then with the *Conjugate Gradient* algorithm. The *neck linker* chain was subsequently surrounded with water molecules (a *Simple Point Charge* model), the whole system was put in a box under *periodic boundary conditions* and optimization algorithms were used again. **MD** simulations were performed on the box. Temperature coupling was handled with the *Berendsen* thermostat (reference temperature $T = 300K$) while pressure coupling was handled with the *Parrinello-Rahman* barostat (reference pressure $p = 100kPa$). A time step Δt was equal to $2 \cdot 10^{-3}ps$. The equilibrium has been reached after accruing *6ns* of **MD** trajectory. An ensemble of *end-to-end* distances of the *neck linker* peptide has been gathered from the whole period of equilibration. A distribution of this ensemble was then created. The mean *end-to-end* distance from that distribution was taken to be the guiding characteristics for the choice of equilibrated structure. I have selected a structure from the time frame near the end of the simulation, for which *end-to-end* distance was sufficiently close to the mean. The explicit water molecules were then replaced by implicit solvent, to lessen computational burden and allow for large number of repetitions. The system was relaxed with a short equilibration routine, while keeping positions of first and last C_α atoms fixed. After that, it was used as a starting point for five ensembles of 10^4 simulation runs. A constant force was applied between a mass centre of the 1st residue and a mass centre of the 19th residue. The ensembles differ because of different values of a stretching force, that were used. These values were: 130, 400, 700, 1000 and $1300 \frac{kJ}{mol \cdot nm}$. The different values are meant to generate different speeds of the stretching and drag the system far from equilibrium. The value of maximum force used translates roughly to 2160 *pN*. The force required to move kinesin motor during *optical trap* experiments has been widely reported as below 10 *pN*[101, 100, 104]. **MD** simulations that focused on testing *neck linker's* properties used force values that surpass 200 *pN*[35, 65]. My goal was to test resilience of the structures against an external influence that may cause arbitrarily high stretching speeds. The maximum limit of force values has been selected after some preliminary test simulations. A forces that test the mechanical properties of various peptides and exceed $1300 \frac{kJ}{mol \cdot nm}$ has been used in **MD** simulations before[34, 64]. The simulations of non-equilibrium dynamics produced five ensembles of 1*ps* trajectories. The final states of these trajectories became starting points of another five ensembles of 10^4 simulations, each lasting 1*ps* ($\Delta t = 2 \times 10^{-3}ps$). These simulations model the the reverse process with regards to the stretching, a relaxation, thus a constant stretching force is not applied. The 4.5.5 version [83] of the **GROMACS** package [8, 69] and the *OPLS-aa* force field [55, 54] have been employed to perform original 10^4 **MD** simulation runs and estimate ΔG values. The 2018 version of the **GROMACS** package with the same force field has been employed to perform additional $2 \cdot 10^3$ runs, which has been used to perform averaged trajectory analysis.

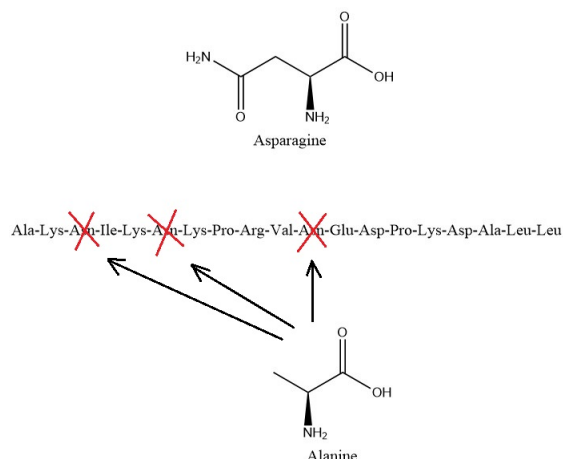


Figure 3.2.1 A scheme of substitutions that led to the *no-Asparagine* chain. Each *Asparagine* residue is being replaced with *Alanine* amino acid.

3.2.2 Poly-alanine

Alanine amino acid is the most simple of them all- its side chain is just an hydrogen atom. That is to say, it doesn't have a side chain. As such, substituting *Alanine* for any residue erases the presence of a side chain in that place in a sequence. For that reasons, I have decided to chose a *poly-alanine* sequence as a reference sequence for the *neck linker*. I created an 18 amino acid long peptide, to which I will henceforth refer as the *Ala-only* peptide or the *Ala-only* chain. All the same steps described above have been taken with regards to the *Ala-only* peptide. One difference being that the equilibration simulation lasted *3ns*. As an aside, it should be noted that the force value about *3600pN* has been used to drive a *deca-alanine* peptide to the stretching limit[34]. That amount of force is well beyond what I applied, so our parameters are well within accepted limit.

3.2.3 Point Mutations

Three additional structures have been created. The original *neck linker* sequence has undergone modifications. I have substituted three amino acid types with *Alanine*, in order to eliminate the influence that their side chains exert upon the structure. The amino acid types selected for substitution are *Asparagine*, *Lysine* and *Proline*. *Asparagine* has been selected because of an amine group's presence in the side chain and its polarity that promotes contacts with water. *Lysine* has been selected because of the exposed amine group in the side chain, while *Proline* has been selected because a heterocyclic ring is present in *Proline*'s side chain. The heterocyclic ring is Pyrrolidine, a ring of carbon atoms with one additional nitrogen atom present. *Proline* is the only residue with a ring in a side chain among amino acids making up *neck linker*. *Proline* can be found either in *cis*- or *trans*- conformations. The removal of these features from the side chain space promises visible impact on the intramolecular contacts. There are three *Asparagine* residues in *neck linker* sequence, four *Lysine* residues and two *Proline* residues. In the *no-Asparagine* peptide all *Asparagine* residues were replaced with *Alanine* (see Fig. 3.2.1). In the *no-Lysine* peptide all *Lysine* residues were replaced with *Alanine* (see Fig. 3.2.2). In the *no-Proline* peptide all *Proline* residues were replaced with *Alanine* (see Fig. 3.2.3).

All the same steps described above for *neck linker* peptide have been taken with regards to each of the 'mutated' peptides. The equilibrium simulations of those three altered linker structures lasted *5ns*, *6ns* and *3ns* for *no-Asparagine*, *no-Lysine* and *no-Proline* chains, respectively.

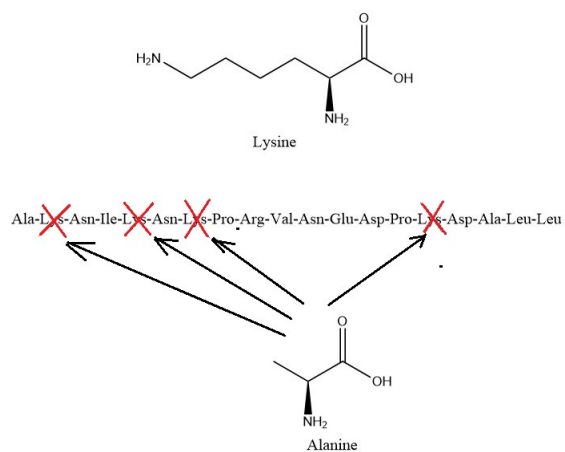


Figure 3.2.2 A scheme of substitutions that led to the *no-Lysine* chain. Each *Lysine* residue is being replaced with *Alanine* amino acid.

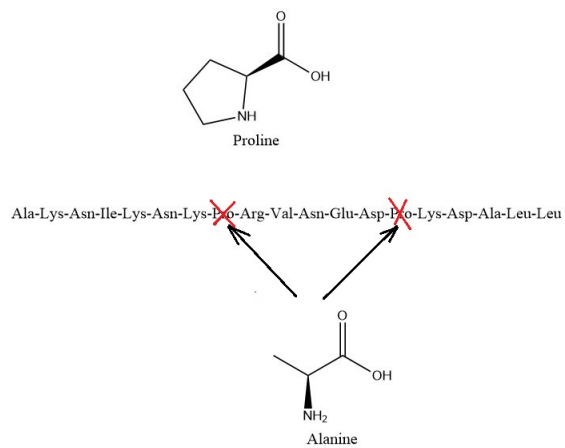


Figure 3.2.3 A scheme of substitutions that led to the *no-Proline* chain. Each *Proline* residue is being replaced with *Alanine* amino acid.

3.3 Results

3.3.1 Equilibration and selection of starting structures for stretching simulations

A starting structure for **MD** simulations should be sampled from a properly equilibrated ensemble. Standard practice involves plotting the energy of the molecule as a function of time. When the function reaches a plateau, the equilibrium is assumed. However, the perception of reaching a plateau is subjective. In the past, when simulation time were shorter, an equilibration run of 100 ps was usually deemed as sufficient[35, 106]. Already back then the assumption was questioned[93, 29]. As the available resources grew, length of equilibration runs grew as well. For a short peptide that is essentially a molecular chain, aside from standard practice of plotting molecule’s energy and **Root Mean Square Displacement (RMSD)** against time, we may consider a distribution of its *end-to-end* distance as a good additional measure of equilibrium. Since the equilibrated structure must be in an energy well that corresponds to a minimum of the potential function, we expect a distribution of its *end-to-end* distance to approach normal distribution as the structure reaches the bottom of the energy well. Since a stretching force will be applied later on along *end-to-end* distance, it is another reason to obtain its well sampled value.

The upper panels of Fig.3.3.1 present the comparison of *end-to-end* distance distributions between *neck linker* and the reference structure, the *Ala-only* peptide. Despite the fact that the *neck linker* was equilibrated for 6 ns and *Ala-only* peptide only for 3 ns , the *Ala-only* peptide’s distribution is well fitted to the Gaussian curve. On the other hand, the *neck linker*’s distribution displays a significantly worse fit. In Fig.3.3.2 we see analogous distributions for ‘mutated’ peptides. The *no-Asparagine* and *no-Proline* peptides reached a good fit to the Gaussian curves in 5 ns and 3 ns respectively. We can assume that the *Ala-only* and the *no-Proline* peptides have reached close to a local minimum of potential energy relatively easily. It took the *no-Asparagine* peptide much more time, but in the end it reached a distribution with a very similar variance σ^2 to that of the *no-Proline* peptide’s distribution. Surprisingly, the *no-Lysine* peptide’s distribution from the whole equilibration run does not fit the Gaussian curve at all, as we can see in the middle panel of Fig.3.3.2. But when we look at Fig.3.3.3, we see that for the subset of 10^4 simulation steps (equivalent to 20 ps) at the end of the equilibration run, the *no-Lysine* peptide’s distribution fits the Gaussian curve extremely well. It could be tempting to assume that if the distribution agrees with the Gaussian curve at any point, it will continue to do so from that point. However, the middle and lower rows of the left column in Fig.3.3.1 clearly illustrate that this assumption would be wrong. Last 10^4 simulation steps of the *neck linker*’s equilibration run showcase a poor fit, while a fit of the preceding 10^4 simulation steps is very good. Additionally, the analogous ranges of simulation steps from the *Ala-only* peptide’s equilibration run showcase a worse fit than the distribution from the whole equilibration run.

These results may suggest that when a structure reaches a local minimum of energy function, long range interactions become muffled. In effect, the structure explores a much narrower set of conformational states, behaving similarly to the Gaussian chain. At some point, thermal fluctuations force the structure away from the bottom of the energy well. With time, if the structure remain close to that particular energy well, the distribution of *end-to-end* distances converges to the normal distribution.

The starting structures for stretching process has been selected from timestamps close to the end of the equilibration runs. The *end-to-end* distance of the starting structure of each peptide matched the relevant mean *end-to-end* distance listed in Table 3.1. In case of *no-Lysine* peptide, the distance listed in Table 3.1 has been calculated for a period of the equilibration run, during which *no-Lysine*’s *end-to-end* distance followed the normal distribution (see Fig.3.3.3 for details). All in all, *no-Lysine*’s starting structure has been drawn from that period of equilibration run.

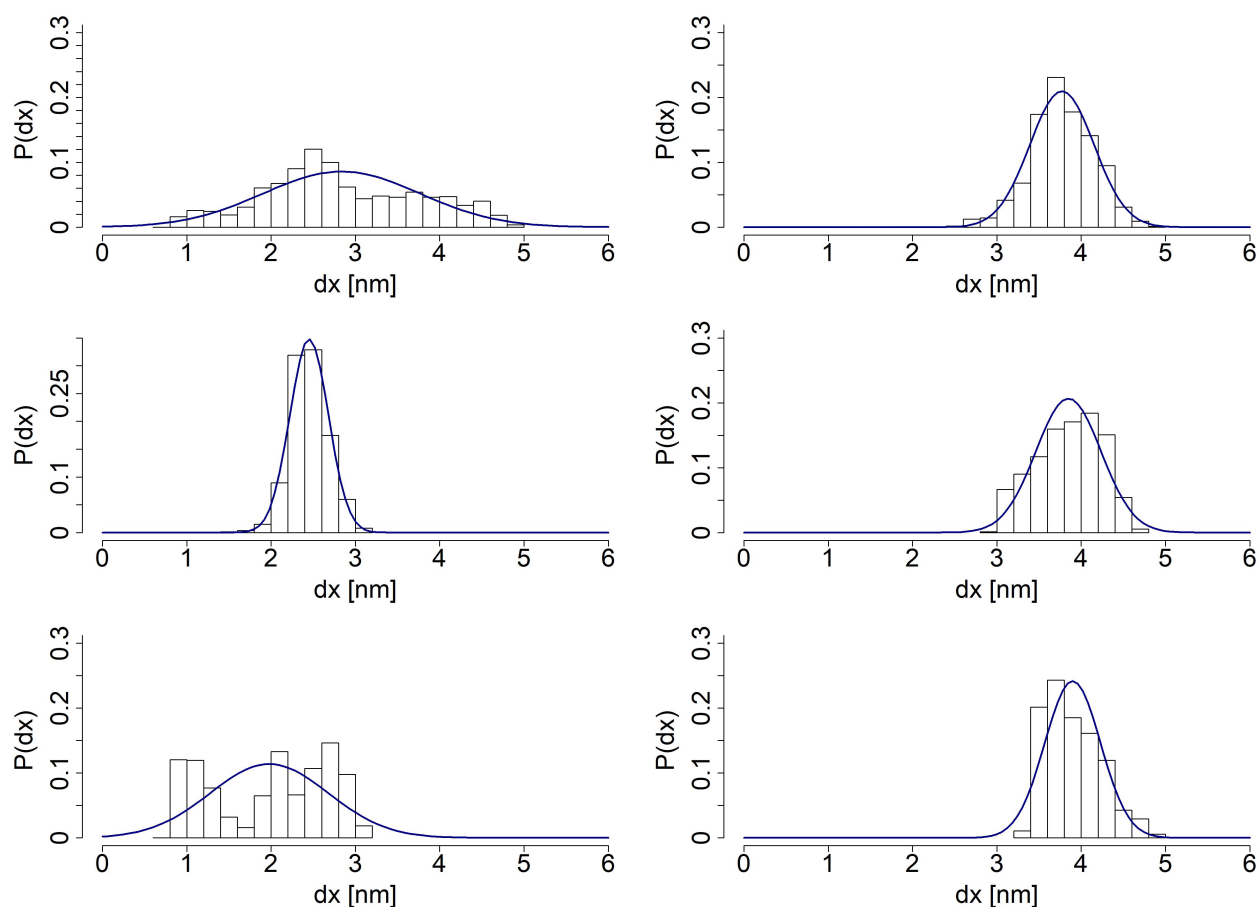


Figure 3.3.1 Left column includes (from top to bottom): the *neck linker's end-to-end* distance distributions from the whole equilibrium simulation, 10^4 recorded steps preceding the last 10^4 and, in the final row, the last 10^4 recorded steps. Right column includes the *Ala-only peptide's end-to-end* distance distributions arranged in an analogous way. Gaussian curves have been fitted to guide an eye, based on the mean and σ^2 of a given empirical distribution derived from the **SMD** simulations.

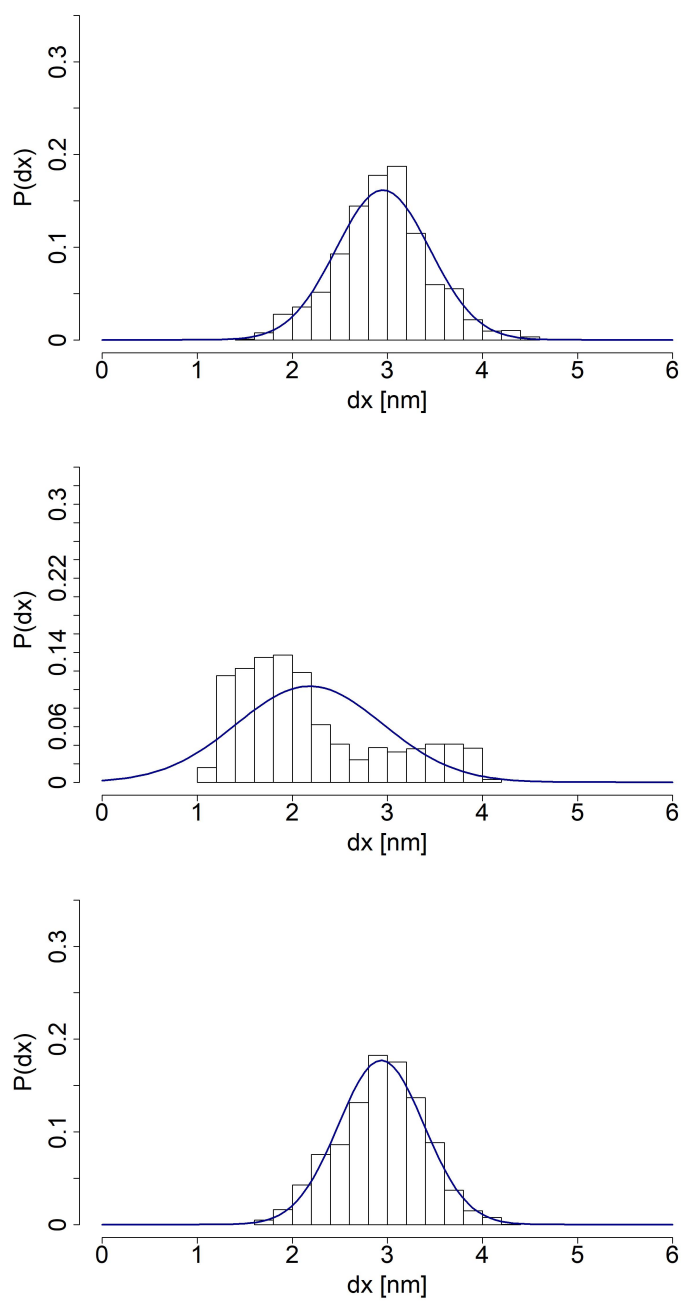


Figure 3.3.2 The *end-to-end* distance distributions for the *no-Asparagine* chain (top), the *no-Lysine* chain (middle) and the *no-Proline* chain (bottom), derived from whole equilibrium simulations. Appropriate Gaussian curves have been fitted to the data, based on derived means and σ^2 of respective distributions.

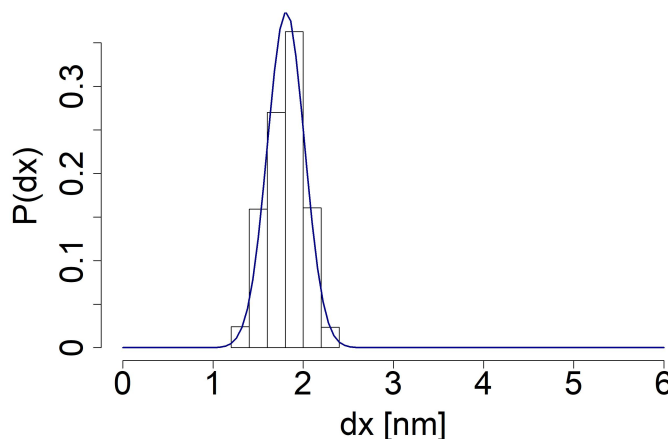


Figure 3.3.3 The *end-to-end* distance distribution of the last 10^4 steps taken from the *no-Lysine*'s peptide equilibration run data. An appropriate Gaussian curve has been fitted to the data, based on its mean and σ^2 .

equilibrated structure	mean <i>end-to-end</i> distance [nm]
<i>Ala-only</i>	3.77 ± 0.38
<i>neck linker</i>	2.27 ± 0.56
<i>no-Asparagine</i>	2.95 ± 0.49
<i>no-Lysine</i>	1.81 ± 0.21
<i>no-Proline</i>	2.94 ± 0.45

Table 3.1: The mean *end-to-end* distances of equilibrated structures with standard deviation. The mean and the deviation of *no-Lysine* structure are derived from last 10^4 steps from the *no-Lysine* peptide's equilibration run (see Fig.3.3.3).

3.3.2 Analysis of trajectories averaged over ensembles of stretched peptides

Basic comparison of *end-to-end* distance's time evolution between *neck linker* and the reference structure, *Ala-only* peptide has been made. They are shown in Fig.3.3.4. As expected, we see that stretching speed varies slightly during simulations, For the *neck linker* the shifts are gradual, while for *Ala-only* peptide the one major shift is instantaneous. Nevertheless, if the speed of *neck linker*'s stretching were to be approximated as having constant value, it would not be far off. The *Ala-only* chain's linear response breaks down more clearly and much sooner than that of the *neck linker* structure. This is in agreement with findings reported in one of my prior publications [71].

The *end-to-end* distance is a property of utmost importance when it comes to biopolymers' description. It must be well known if one wants to perform **AFM** experiments or recreate them *in silico*. With that being said, an *end-to-end* distance by necessity subsumes the information present in the structure about specifics of intramolecular interactions. One could try to circumvent this issue and squeeze more information out of intramolecular distance changes. Obtaining the information would make it possible to assess which parts of the chain drag behind during the stretching process of the *neck linker* and how it differs for *Ala-only* peptide, the reference structure. Pair distances between the 4th residue and residues of amino acids, previously selected for substitution, have been measured. *Isoleucine* is an amino acid that is among best conserved elements at the 4th

position in a *neck linker* sequence across numerous kinesin families[35]. *Alanine* is obviously the 4th residue in the *Ala-only* peptide. In Fig.3.3.5, there is a list of residues for which pair distances have been measured. The significantly faster growth of pair distances in the *Ala-only* peptide is well within expectations. The lack of side chains must be an important factor in there not being meaningful intramolecular contacts. Displacement patterns between *Isoleucine* at 4th position and subsequent *Alanine* residues differ significantly from those observed for the *neck linker*. The differences in pair distances in the *Ala-only* peptide become apparent after 0.4ps at the latest. While for the *neck linker*, there were no meaningful differences in pair distances in the *neck linker*'s structure up until 0.6ps. After that point the differences become noticeable, which means that a number of intramolecular contacts dropped.

Looking more closely at Fig.3.3.5 we start with those residues that showcase proximal contact with *Isoleucine*. The 2nd, 3rd, 5th and 6th residue can be considered here. For these residues, pair distances change little during the stretching process, regardless of the structure. Of course, this is reasonable. There must be a relatively low boundary on a pair distance, resulting from the length of covalent bonds, when pair members are in immediate vicinity of each other. Residues separated by more than 3 other residues, sooner or later, experience a rapid growth of pair distance. The pair distances defined for 8th, 11th, 14th and 15th residues predictably end up reaching highest values for both peptides. However, after 0.6ps mark pair distance value are more evenly spread within the *neck linker* structure (the 7th residue, *Lysine*, joining the aforementioned four residues in "large distance" category), while in the *Ala-only* peptide's case these four continue to monopolize pair distance growth. It is possible that this *Lysine* residue helps in mediating the strain from stretching towards the N-terminus and its absence in the reference structure causes a dysfunction. Despite the fact that pair distances in the *Ala-only* chain reach maximum values almost twice as large as maximum values reached in the *neck linker* structure, we see in Fig.3.3.4 that a *end-to-end* distance change is much larger for the *neck linker* than for the *Ala-only* chain. Which means that average pair distance values must be higher in the *neck linker* structure. *Proline* occupies positions in both the 8th and 14th residue. When a *Proline* residue is in a (*cis*-) conformation, *Proline* has been reported to have a negative impact on the chain's flexibility[73, 35]. In fact, *in silico* studies into mechanical properties of the *neck linker* domain from sequence analysis [35] show that the *Proline*, that constitutes the 14th residue, is found usually in *cis*- conformation and that effectively shortens the peptide. Additionally, *cis-trans* switching of a conserved *Proline* residues under the influence of a large force, accounts for variations in resulting force-extension profiles. It agrees with experimental results[73], which indicate that the proline's isomerization influences duration of biological processes dependent on protein folding and can be seen as a timer of these biological processes.

In Fig. 3.3.6 we can observe changes in potential energy for all simulated peptide ensembles. Interestingly, as simulation time passes in most cases potential energy decreases or remains at the same level, regardless of the fact that we are introducing a disruptive force and taking system away from equilibrium. That would mean that in most instances peptides traverse attractive, low-energy conformations. Interestingly, in all instances highest applied force leads to lowest Potential energy. As we focus on individual energy profiles in Fig. 3.3.6, we see that *Ala-only* series seem to diverge into different regimes at early steps, but as the simulations continues, they remain very stable. At low forces: $130 \text{ kJ} \cdot \text{mol}^{-1} \cdot \text{nm}^{-1}$ (216pN) and $400 \text{ kJ} \cdot \text{mol}^{-1} \cdot \text{nm}^{-1}$ changes in *Ala-only* series are so minuscule, it could be erroneously assumed that the molecule stays in equilibrium, which we know is not the case. However, while Potential energy appears stable, it is actually much higher than for the rest of the peptides. In comparison potential energy values of *neck linker* ensemble differ significantly. First steps show energy values relatively close to each other. While energy distance between top and bottom series of *Ala-only* peptide ensembles is around $4000 \text{ kJ} \cdot \text{mol}^{-1}$, for *neck linker* ensembles it is below $1500 \text{ kJ} \cdot \text{mol}^{-1}$. Additionally, *neck linker* peptide is the only one for whom Potential energy series from separate ensembles cross. As we look at energy fluctuation patterns of the rest of the peptide ensembles, it seems they are more similar to the those of *Ala-only* peptide. *no-Asparagine* peptide's energy ensembles traverse in separate energy regimes, in similar fashion to these of *Ala-only* peptide. Same for *no-Proline* peptide ensembles. It is worth noting however that

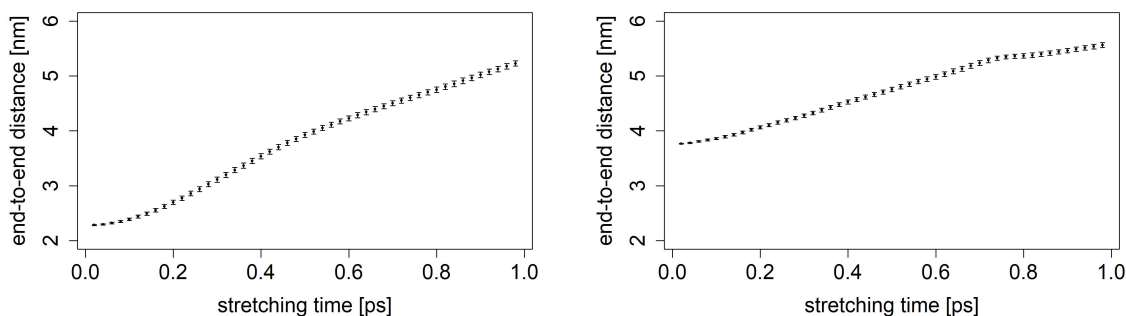


Figure 3.3.4 *End-to-end* distances as functions of time, averaged over 10^4 **MD** runs, each of duration $t = 1$ ps. The left plot depicts a change in the *neck linker's* end-to-end distance. The right plot displays analogous findings for the *Ala-only* chain.

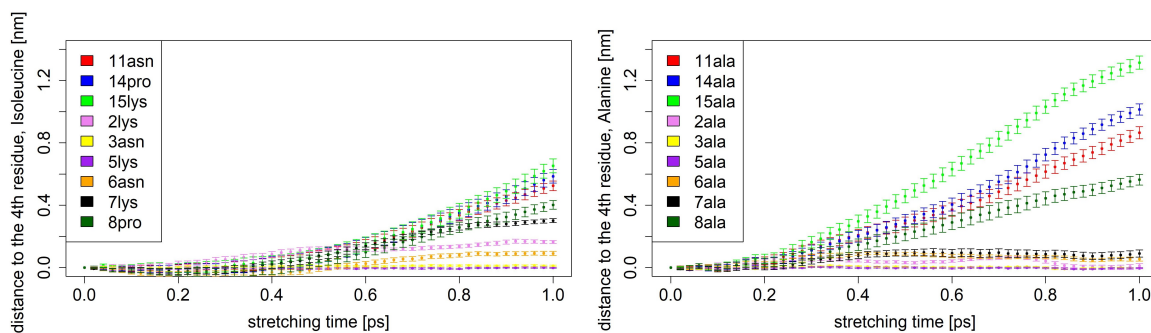


Figure 3.3.5 Distances (extensions) between residues as functions of time, averaged over 10^4 **MD** runs, each of duration $t = 1$ ps. The left plot depicts different rates of distance changes in the *neck linker's* chain between the 4th residue and selected residues, listed in the inset. The right plot displays analogous findings for the *Ala-only* chain.

variance in energy estimation is much higher for other peptides, when compared to *Ala-only*.

One curious case is *no-Lysine* peptide. Even though its energy series seem to show tendencies much like the others, with similar variance of energy estimation, its series showing the response to the constant force $1300 \text{ kJ} \cdot \text{mol}^{-1} \cdot \text{nm}^{-1}$ experiences three spikes of variance. Correspondingly, as spikes occur, the averaged potential energy of the ensemble experiences slight upticks, despite an overall downward trend. It would seem that each spike is tied to a saddle point in conformational hyperspace that some members of the ensemble stumbled upon. This seems to be confirmed by changes in *Lennard-Jones* potential of the relevant *no-Lysine* peptide ensemble, as seen in Fig.3.3.7, with one difference - the 3rd spike does not occur in LJ energy.

Looking further at Fig.3.3.7, we see that *Lennard-Jones* energy in general follows the trend seen in Potential energy in that it decreases with growth of the stretching force. LJ energy values seem to stabilize for the ensembles of *Ala-only* peptide, after experiencing fluctuations the beginning 0.2 ps . *neck linker* series are the only one to show clear growth in the beginning 0.2 ps . The *neck linker* ensembles also move in regime of higher LJ energy values, while *Ala-only* peptide ensembles are in lowest regime and ensembles of *no-Asparagine*, *no-Proline* and *no-Lysine* in the middle.

We have used **RMSD** and ρ to assess the magnitude of change in stretched structures in comparison to initial structures. In Fig.3.3.8 we look at how these changes depend on stretching force for one type of peptide and in Fig.3.3.9 we look at how the peptide ensembles differ at each stretching force level. In Fig.3.3.8 we immediately notice that for all peptides both **RMSD** and ρ grow in time. The growth is directly tied to the magnitude of the stretching force present in the system, as expected. It is assumed that a threshold, beyond which we may consider compared structures to be dissimilar, is $0.2 - 0.3 \text{ nm}$ in case of **RMSD** and $0.3 - 5$ in case of ρ [74]. If we take upper limit for **RMSD** and lower limit for ρ (that is conveniently 0.3 nm and 0.3 respectively), we see that for each peptide **RMSD** series reach the threshold before the corresponding ρ series. Comparing just at highest stretching force level, *Ala-only* and *neck linker* series reach **RMSD** threshold below a time stamp of 0.5 ps , while ρ threshold is reached after that time stamp. Similar situation occurs for *no-Asparagine* and *no-Proline* peptides, but in this instance the time stamp is 0.6 ps . For *no-Lysine* peptide the time stamp is 0.7 ps . That clearly shows that **RMSD** is more sensitive to the changes in conformation for short peptides.

Another instance, where the difference in sensitivity can be noticed, is on display in Fig.3.3.9. While in the presence of stretching force of $130 \text{ kJ} \cdot \text{mol}^{-1} \cdot \text{nm}^{-1}$, both measures do not consider any of the peptides to be dissimilar, as stretching force value increases, ρ has difficulty differentiating between various structures and it improves only slightly as the stretching force rises. **RMSD** on the other hand already differentiates ensembles significantly better when middle level force is applied and improves from then on. It might be connected to the fact that ρ has to fit all possible distance values into $0 - 2$ range. Crucially, *no-Lysine* peptide ensembles visibly fall off in **RMSD** plots, hinting at *no-Lysine* peptide being significantly more resistant to deformation than the other peptides, while ρ is not making that clear at all. If we additionally account for the fact the **RMSD** results give us concrete values that correspond with actual distances in the system, we can conclude that **RMSD** is a more useful measure when analyzing short chains.

While we have gained some general insight into the peptides' behaviour looking at **RMSD** and ρ , radius of gyration R_g gives us more detailed information since it measures mass distribution around the center of mass. Radius of gyration R_g is expected to rise in value the more a short peptide is stretched and the more similar its shape has become to a shape of a rod. In Fig.3.3.10 we see how R_g changes for each peptide under different stretching forces.

One thing that becomes immediately clear is that, at the lowest force level, the ensembles of *neck linker* and other mutated peptides do not experience a monotonic increase like the *Ala-only* peptide ensemble does. It would mean that *Ala-only* peptide offers hardly any resistance to the stretching and unfolds smoothly. *Neck linker* and other mutated peptides experience a momentary drop in R_g values, which would mean their structure is consolidated around the center of mass for a time. In each case R_g eventually starts increasing, except the *no-Proline* peptide ensemble, where the decreasing trend is upheld for the span of the whole simulation. The

initial R_g values are different for ensembles of each peptide, which is expected since each peptide has been equilibrated independently. The R_g increase is tied to the stretching force increase. The *Ala-only* peptide ensembles show strongest response to stretching forces, while *no-Lysine* peptide is the most resistant. As we compare R_g at each stretching force level (see Fig.3.3.11) we notice that *no-Proline* peptide series at all levels follow closely *neck linker* series. This would suggest that eliminating Proline residues brings the least change to neck linker functionality, but as I have observed and commented on above (when discussing Fig.3.3.6, Fig.3.3.7 and Fig.3.3.9) and will do so shortly below (a discussion of Fig.3.3.12 and Fig.3.3.13), the relation is not really that straightforward. It is worth noting that *no-Asparagine* series in Fig.3.3.11, even though starting from lower initial R_g values, catch up to *no-Proline* and *neck linker* series in the latter part of simulation runtime. How quickly *no-Asparagine* series catch up depends on stretching force. As expected, *no-Lysine* series stay separate from the rest, barely reaching initial R_g values of the other peptides even when being stretched by the force of $1300 \text{ kJ} \cdot \text{mol}^{-1} \cdot \text{nm}^{-1}$.

Aside from knowing R_g , knowledge of how many hydrogen bonds there are within the peptide structures would also provide some hints regarding similarities in conformational changes between different peptides. I have set hydrogen bonds' parameters as follows: the minimum distance between donor and acceptor to 0.35 nm and minimum angle between donor, hydrogen and acceptor to 30° . These are identical to default settings for hydrogen bond calculation in **GROMACS** 2018. Standard deviation values for each mean hydrogen bond for each averaged hydrogen bond number have been ranging roughly from 0.5 during initial simulation steps up to 2.0 near the end of simulations at highest level of stretching force. Comparable range of hydrogen bond number variance has been previously observed for short peptides[80]. Considering that average number of hydrogen bonds for our peptides hardly goes beyond 6, it would not be productive to consider our results in Fig.3.3.12 and Fig.3.3.13 as precise predictions of actual hydrogen bond number. Rather, I treat the patterns of change in hydrogen bonds' number as a piece of information giving us some insight into inner workings of the system. In Fig.3.3.12 a number of overall hydrogen bonds and a number of hydrogen bonds within 1 nm of $10^{\text{th}}C_\alpha$ atom for each peptide are compared. It can be observed that the difference between them is small for *Ala-only*, *neck linker* and *no-Proline* peptides. For each of these peptides the pattern of change in hydrogen bonds' number is also very consistent between the overall structure and 1 nm radius around $10^{\text{th}}C_\alpha$ atom. This tendency is however not upheld when it comes to *no-Lysine* and *no-Asparagine* peptides. In their case, there is significant difference in number and in pattern. For the three peptides mentioned first, after initial period (about 0.4 ps for *Ala-only*, 0.6 ps for *neck linker* and 0.8 ps for *no-Proline* peptide), series diverge in accordance to the level of stretching force, with number of hydrogen bonds subsequently decreasing as the magnitude of the stretching force rises. Surprisingly, series of *no-Asparagine* ensembles fail to replicate this straightforward dependence on stretching force and it is the only instance in our analysis where this kind of situation occurs. It seems that main culprit for that behaviour is *no-Asparagine* ensemble that interacts with the stretching force of $1300 \text{ kJ} \cdot \text{mol}^{-1} \cdot \text{nm}^{-1}$. For the series of that ensemble, while the overall number of hydrogen bonds decreases with time as expected, the number of hydrogen bonds around $10^{\text{th}}C_\alpha$ atom stays roughly constant. In this *no-Asparagine* peptide is actually similar to *neck linker* peptide, the difference being that *neck linker* ensembles show greater increase of hydrogen bonds' number at lower force levels and because of that for *neck linker* peptide the typical pattern of stretching force dependence is preserved at the end of the simulation runs.

As we analyze patterns at each force level in Fig.3.3.13, two key features become apparent: initial overall hydrogen bonds' number being much higher in case of *no-Lysine* peptide compared to the rest and a number of hydrogen bonds increasing rapidly mid-run in case of *no-Proline* peptide. It is worth noting that the patterns of *neck linker* and *no-Proline* peptides are completely different despite the striking similarity of their R_g values' evolution shown previously. As for the first feature mentioned above, a higher initial number of hydrogen bonds does not reoccur for *no-Lysine* ensembles around $10^{\text{th}}C_\alpha$ atom. That would mean there are conformations *no-Lysine* peptide is prone to assuming, which involve more folding on the ends of the peptide instead of folding in the middle.

We complete our trajectory analysis by looking at **RMSF** profiles for each peptide in Fig.3.3.14. Once again we observe that a measure, in this case **RMSF**, is tied to the stretching force in a way, where a stretching force increase results in **RMSF** increase. This is true for all peptides and aligns well with our expectations. On the other hand, **RMSF** distribution within amino acid chain does differ between the peptides. Going on atom-per-atom basis, for *neck linker*, *no-Asparagine* and *no-Proline* peptides **RMSF** values are significantly higher at both ends of amino acid chains when compared to the **RMSF** values in the middle of the chains. The difference becomes more pronounced as the stretching force increases. In case of *neck linker* peptide ensemble acted upon by the stretching force of $1300 \text{ kJ} \cdot \text{mol}^{-1} \cdot \text{nm}^{-1}$, **RMSF** values almost reach 0.8 nm at N-terminus and around 0.6 nm at C-terminus, while the values in the middle of the chain (roughly between 100^{th} and 200^{th} atom) do not go past 0.2 nm . The value of 0.2 nm can be considered a threshold of significant change in distance and structure, as had been already mentioned while discussing changes in **RMSD** and ρ in the text above.

The *no-Asparagine* and *no-Proline* ensembles acted upon by the stretching force of $1300 \text{ kJ} \cdot \text{mol}^{-1} \cdot \text{nm}^{-1}$ present more symmetrical **RMSF** distributions with **RMSF** values staying below 0.6 nm at both ends. When it comes to *Ala-only* peptide the distribution is even more flat with **RMSF** values below 0.2 nm only for atoms in the narrow range in the middle of the chain (roughly from 75^{th} atom to 100^{th} atom). On the other end of the spectrum, **RMSF** values of *Ala-only* peptide ensembles cross the threshold of 0.4 nm and go no further than around 0.5 nm only for few atoms close to the N-terminus of the chain. The *no-Lysine* peptide is least responsive to stretching force increase among the peptides with the 0.4 nm **RMSF** threshold not being reached at all and the 0.2 nm **RMSF** threshold being clearly exceeded only at the N-terminal region, roughly in range from 1^{st} to 50^{th} atom. It should be noted that overall number of atoms that compose *Ala-only* and *no-Lysine* peptides is much lower than the overall number of atoms that compose previous three peptides. Difference in overall atom number does not however impact the length of amino acid chains, as the atoms in question are situated in side chains, length of peptide backbone is unaffected. In general, it seems that for all tested peptides N-terminus was more prone to fluctuation than C-terminus. It is a reminder that stretching force, that had been applied to C-terminus, took away some fluctuation freedom from C-terminal region.

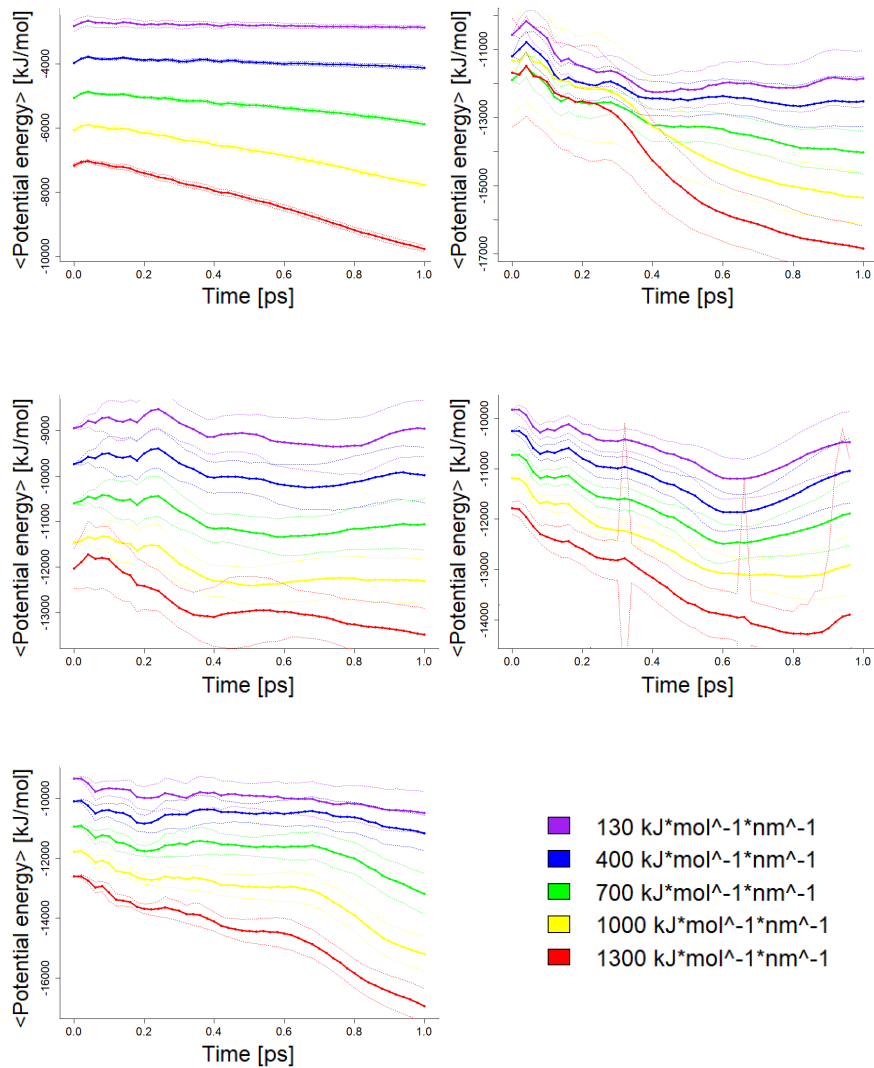


Figure 3.3.6 The change in potential energy over time. Each panel corresponds to each stretched peptide. The legend lists constant forces used to stimulate stretching. The panels will be described from left to right, from up to down. The upper left panel showcases results acquired for *Ala-only* chain. The upper right panel showcases results acquired for *Neck linker*. The middle left panel showcases results acquired for *no-Asparagine* chain. The middle right panel showcases results acquired for *no-Lysine* chain. The bottom panel showcases results acquired for *no-Proline* chain. The data has been acquired in 2×10^3 runs in **GROMACS** 2018. The estimation of potential energy has been achieved by calculating values for each run and then averaging over the ensemble. Dotted lines signify standard deviations.

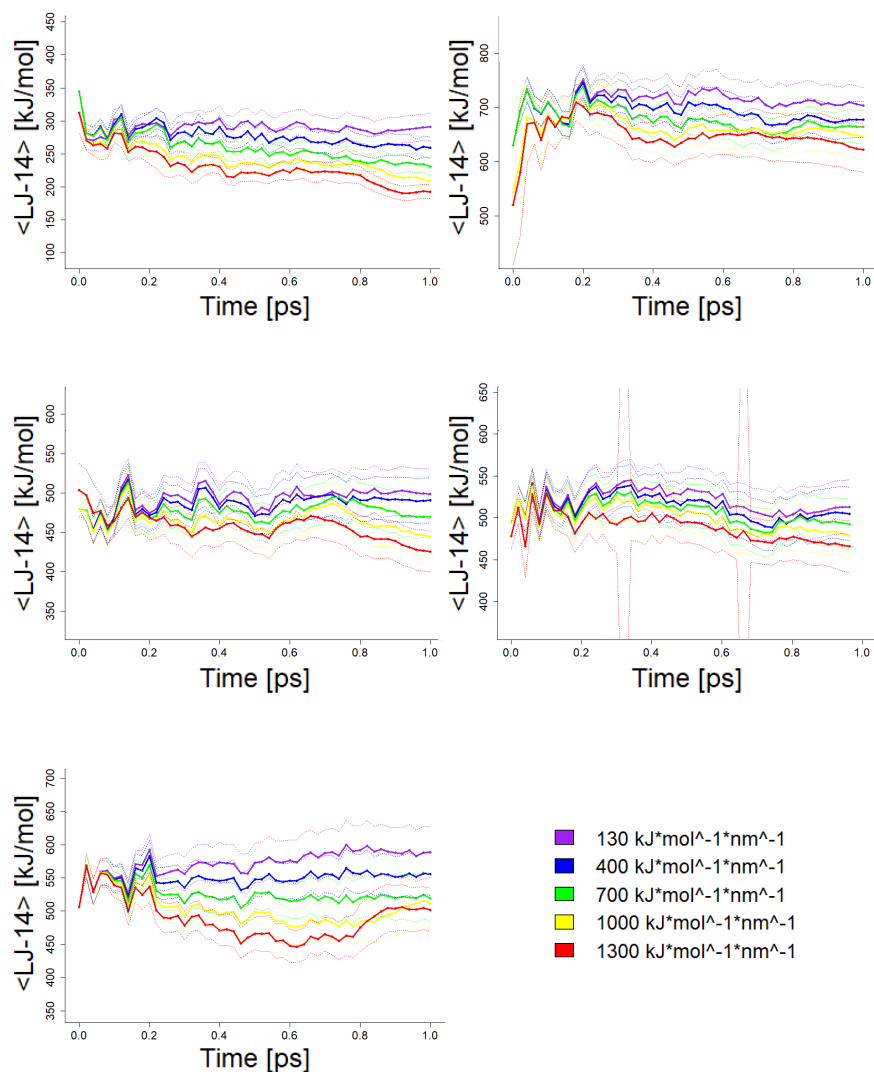


Figure 3.3.7 The change in Lennard-Jones potential over time. Each panel corresponds to each stretched peptide. The legend lists constant forces used to stimulate stretching. The panels will be described from left to right, from up to down. The upper left panel showcases results acquired for *Ala-only* chain. The upper right panel showcases results acquired for *Neck linker*. The middle left panel showcases results acquired for *no-Asparagine* chain. The middle right panel showcases results acquired for *no-Lysine* chain. The bottom panel showcases results acquired for *no-Proline* chain. The data has been acquired in $2 \cdot 10^3$ runs in **GROMACS** 2018. The estimation of Lennard-Jones potential has been achieved by calculating values for each run and then averaging over the ensemble. Dotted lines signify standard deviations.

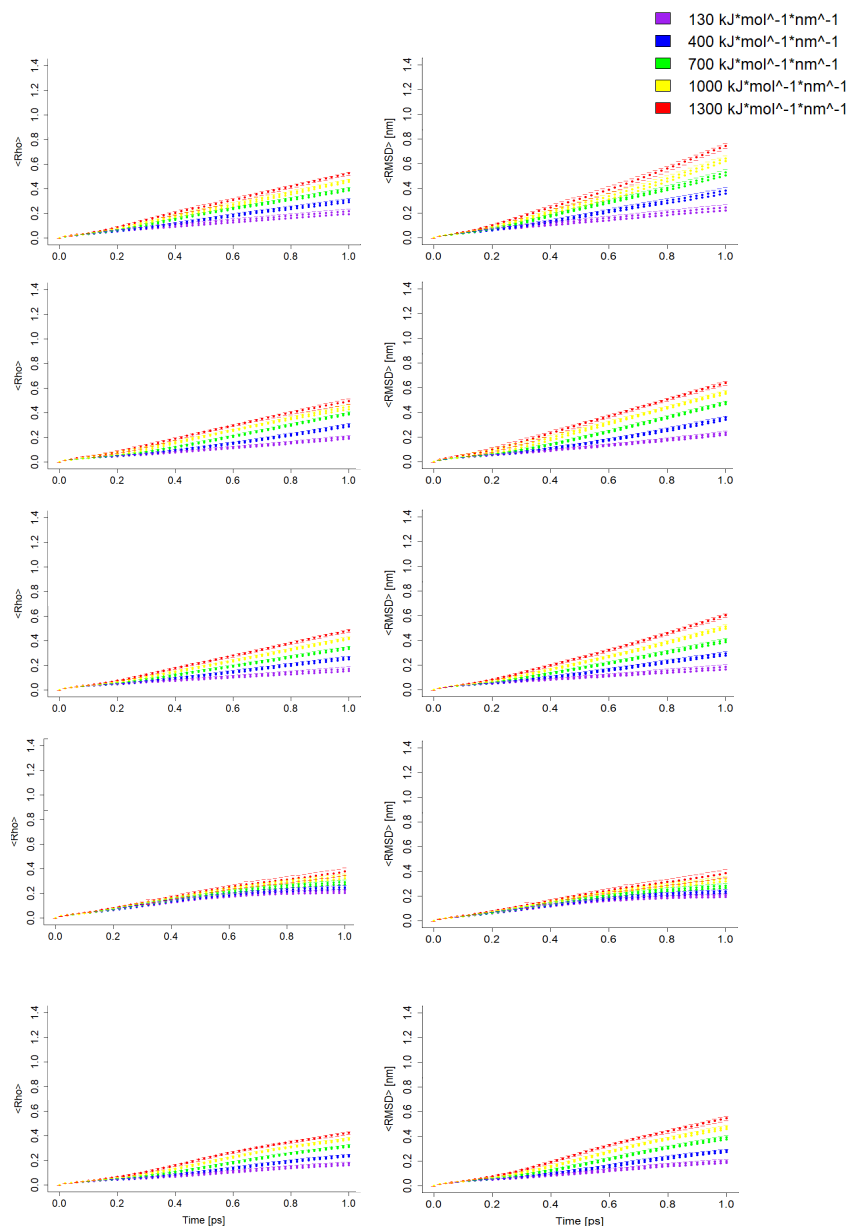


Figure 3.3.8 The change in ρ and **RMSD** over time. The legend lists constant forces used to stimulate stretching. Left column displays ρ values. Right column displays **RMSD** values. Each row is assigned to each stretched peptide: 1st row: *Ala-only*; 2nd row: *neck linker*; 3rd row: *no-Asparagine*; 4th row: *no-Lysine*; 5th row: *no-Proline*. Calculations were based on $2 \cdot 10^3$ runs in **GROMACS** 2018. The estimation of ρ and **RMSD** has been achieved by calculating values for each run and then averaging over the ensemble. Dotted lines signify standard deviations.

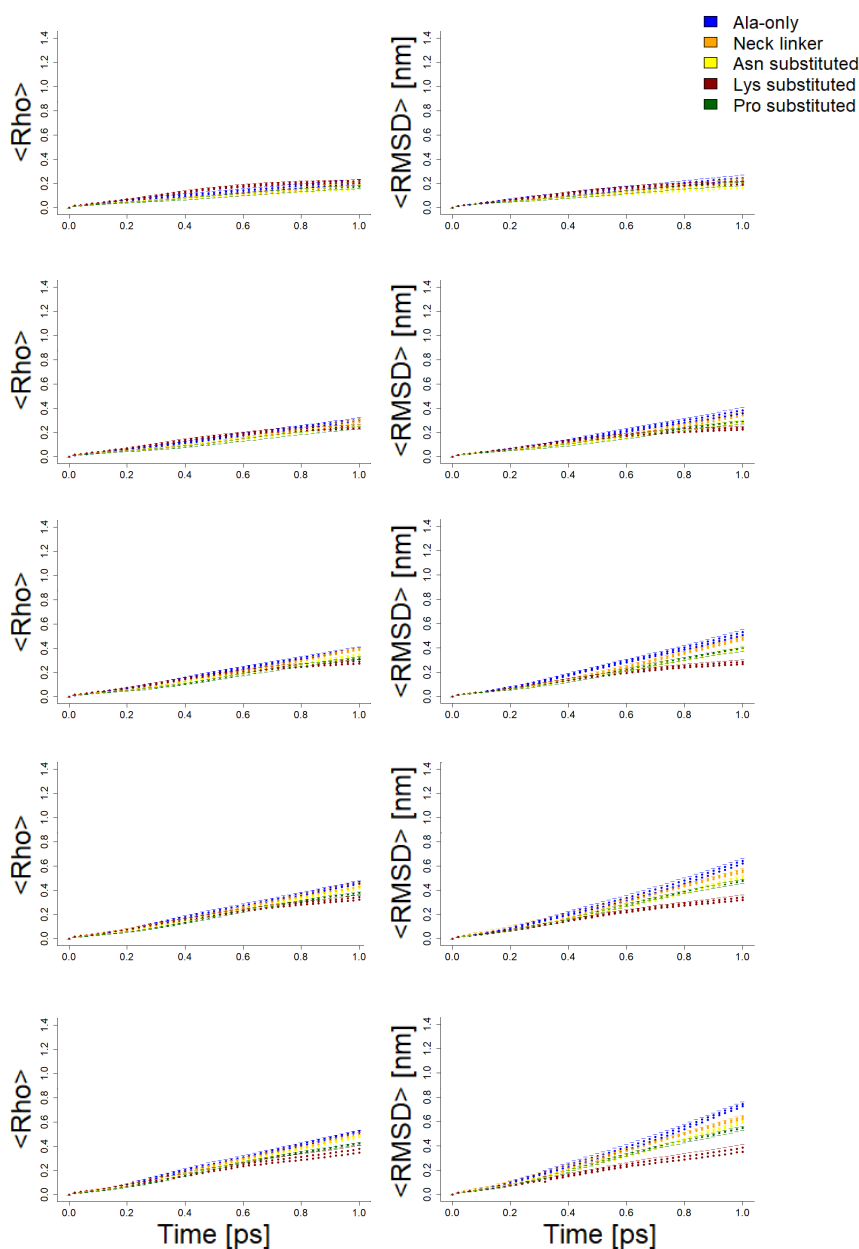


Figure 3.3.9 The change in ρ and **RMSD** over time. The legend lists simulated peptides. Left column displays ρ values. Right column displays **RMSD** values. Peptides were being stretched by constant forces of following values: 1st row: 130; 2nd row: 400; 3rd row: 700; 4th row: 1000; 5th row: 1300. Force unit was $1 \text{ kJ} \cdot \text{mol}^{-1} \cdot \text{nm}^{-1}$. Calculations were based on 2×10^3 runs in **GROMACS** 2018. The estimation of ρ and **RMSD** has been achieved by calculating values for each run and then averaging over the ensemble. Dotted lines signify standard deviations.

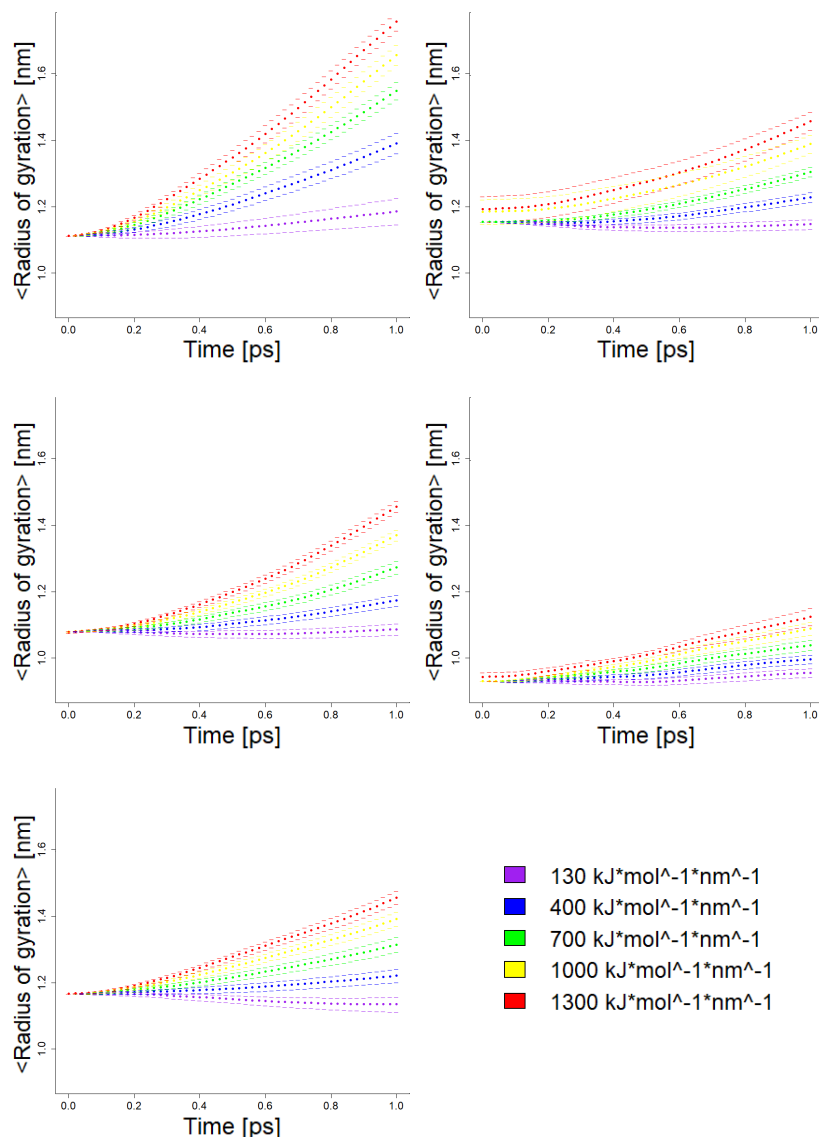


Figure 3.3.10 The change in radius of gyration over time. Each panel corresponds to each stretched peptide. The legend lists constant forces used to stimulate stretching. The panels will be described from left to right, from up to down. The upper left panel showcases results acquired for *Ala-only* chain. The upper right panel showcases results acquired for *Neck linker*. The middle left panel showcases results acquired for *no-Asparagine* chain. The middle right panel showcases results acquired for *no-Lysine* chain. The bottom panel showcases results acquired for *no-Proline* chain. The data has been acquired in $2 \cdot 10^3$ runs in **GROMACS** 2018. The estimation of radius of gyration has been achieved by calculating values for each run and then averaging over the ensemble. Dotted lines signify standard deviations.

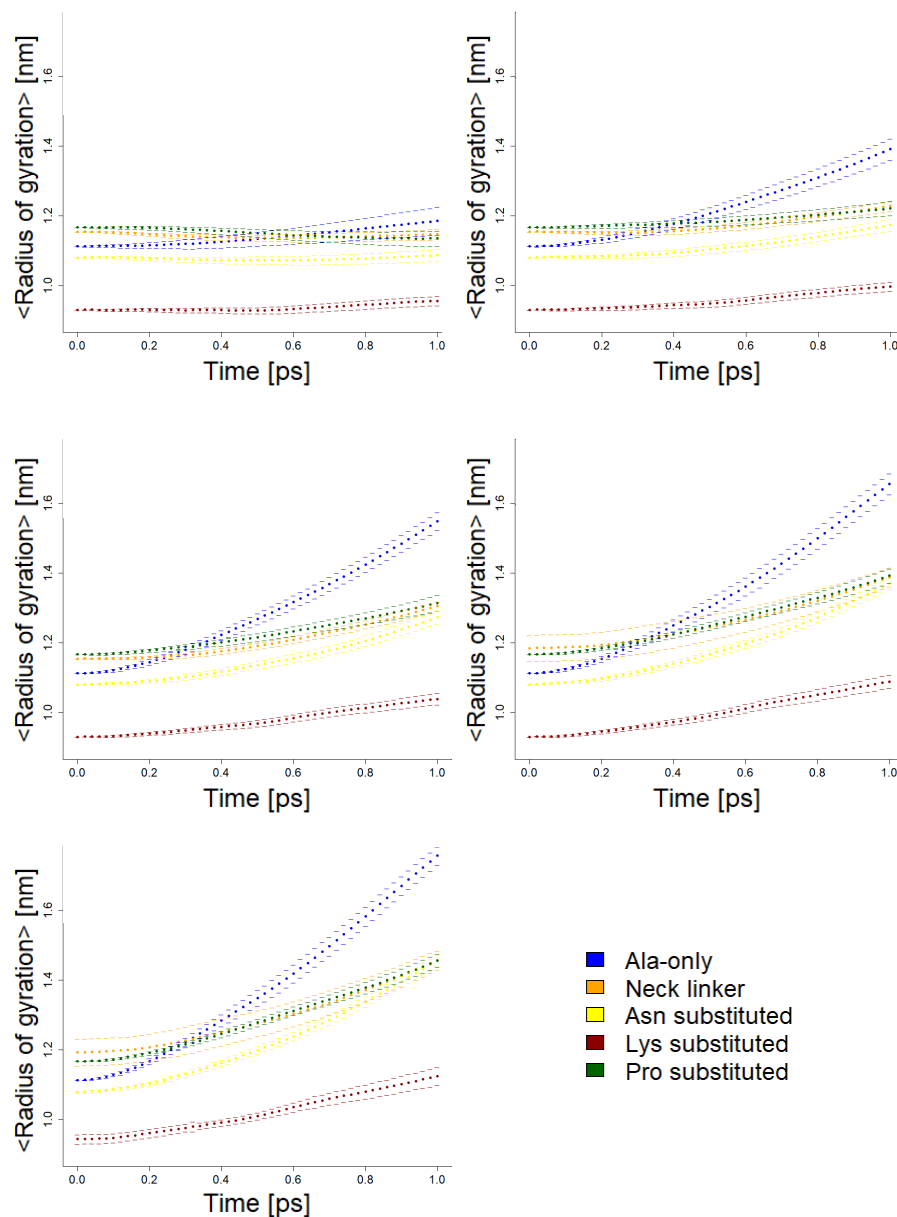


Figure 3.3.11 The change in radius of gyration over time. Each panel corresponds to one of the forces used for stretching. The legend lists peptides that were being stretched. The panels will be described from left to right, from up to down. The upper left panel showcases results acquired under the force of $130 \text{ kJ} \cdot \text{mol}^{-1} \cdot \text{nm}^{-1}$. The upper right panel showcases results acquired under the force of $400 \text{ kJ} \cdot \text{mol}^{-1} \cdot \text{nm}^{-1}$. The middle left panel showcases results acquired under the force of $700 \text{ kJ} \cdot \text{mol}^{-1} \cdot \text{nm}^{-1}$. The middle right panel showcases results acquired under the force of $1000 \text{ kJ} \cdot \text{mol}^{-1} \cdot \text{nm}^{-1}$. The bottom panel showcases results acquired under the force of $1300 \text{ kJ} \cdot \text{mol}^{-1} \cdot \text{nm}^{-1}$. The data has been acquired in $2 \cdot 10^3$ runs in **GROMACS** 2018. The estimation of radius of gyration potential has been achieved by calculating values for each run and then averaging over the ensemble. Dotted lines signify standard deviations.

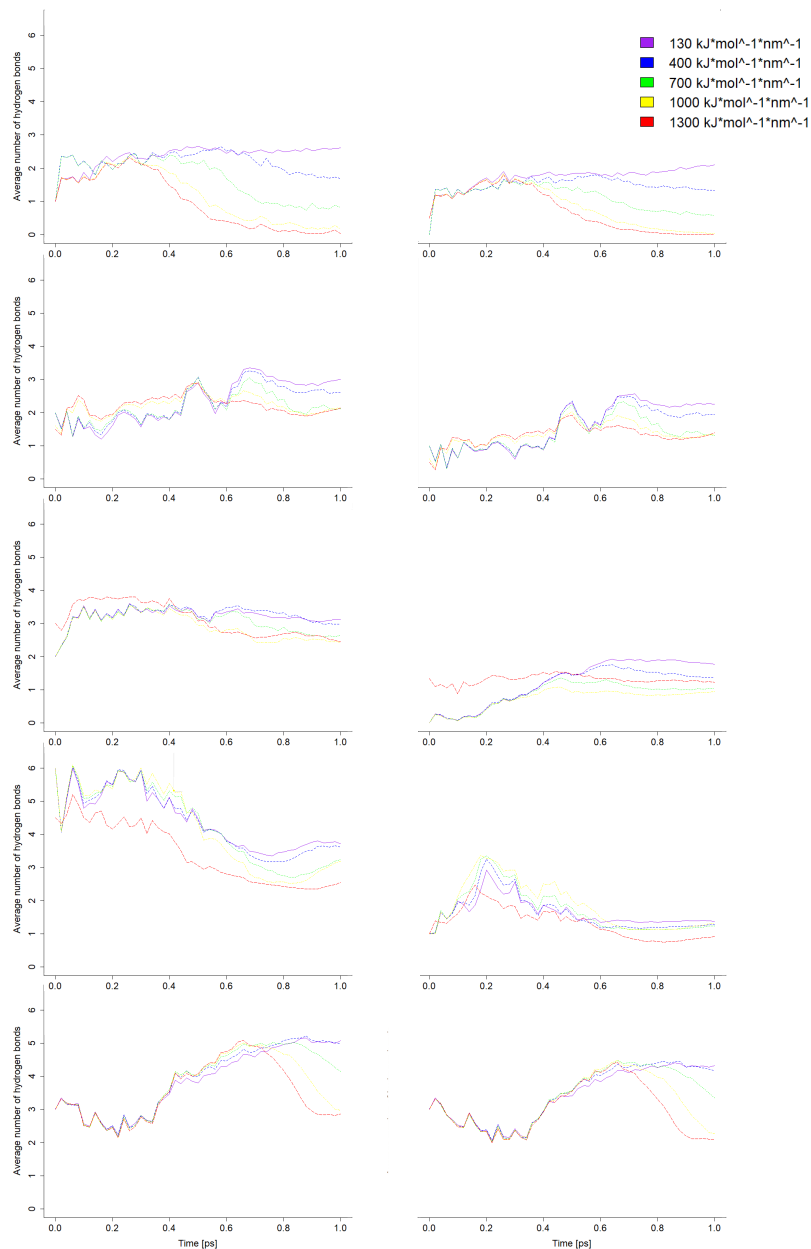


Figure 3.3.12 The change in the average number of hydrogen bonds within peptide structure over time. The legend lists constant forces used to stimulate stretching. Left column displays average number of all hydrogen bonds in a peptide. Right column displays average number of hydrogen bonds within 1nm of C α of 10th residue. Each row is assigned to each stretched peptide: 1st row - results for *Ala-only* chain; 2nd row - results for *neck linker*; 3rd row - results for *no-Asparagine* chain; 4th row - results for *no-Lysine* chain; 5th row - results for *no-Proline* chain. Calculations has were based on $2 \cdot 10^3$ runs in **GROMACS** 2018. The estimation of average numbers of hydrogen bonds has been achieved by calculating values for each run and then averaging over the ensemble.

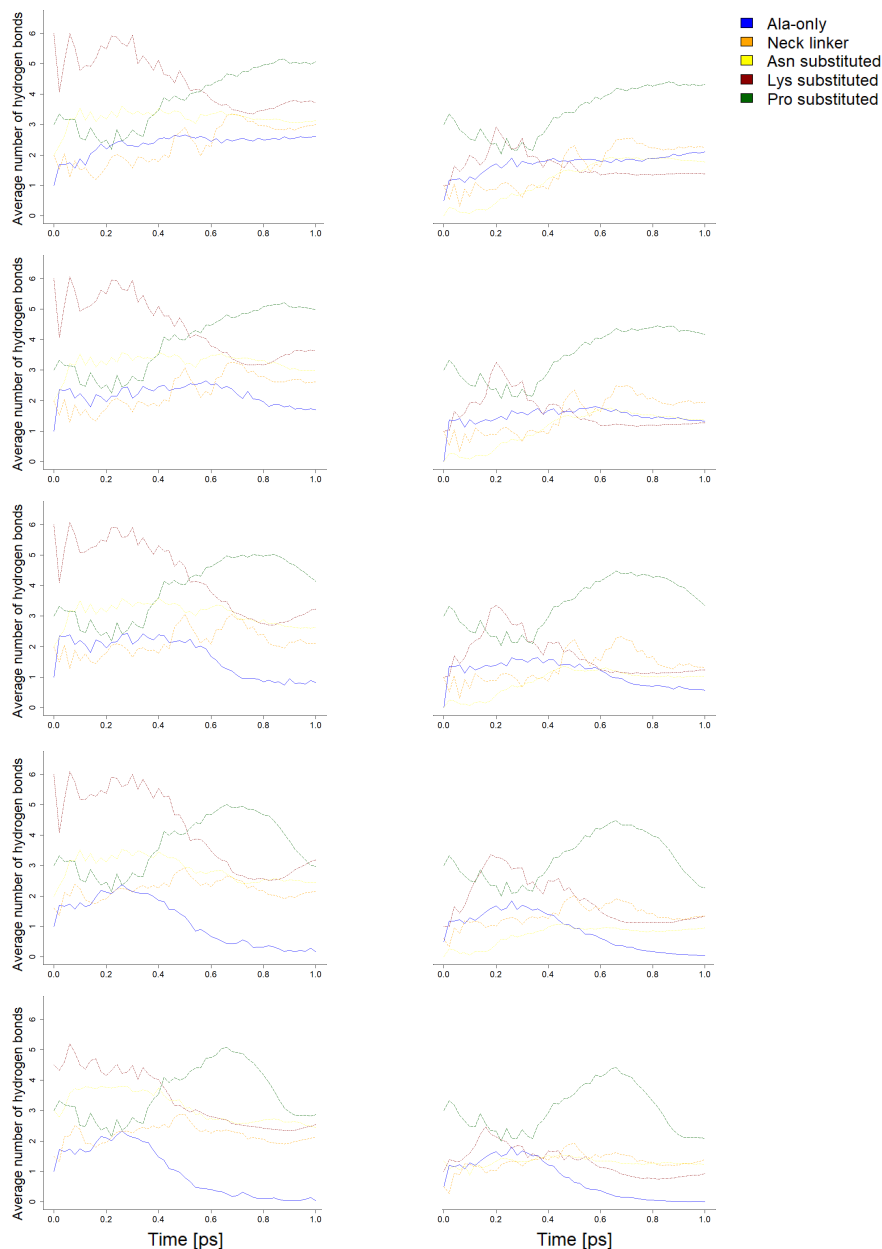


Figure 3.3.13 The change in the average number of hydrogen bonds within peptide structure over time. The legend lists simulated peptides. Left column displays average number of all hydrogen bonds in a peptide. Right column displays average number of hydrogen bonds within 1 nm of C_{α} of 10th residue. Peptides were being stretched by constant forces of following values: 1st row: 130; 2nd row: 400; 3rd row: 700; 4th row: 1000; 5th row: 1300. Force unit was $1 \text{ kJ} \cdot \text{mol}^{-1} \cdot \text{nm}^{-1}$. Calculations were based on 2×10^3 runs in **GROMACS** 2018. The estimation of average numbers of hydrogen bonds has been achieved by calculating values for each run and then averaging over the ensemble.

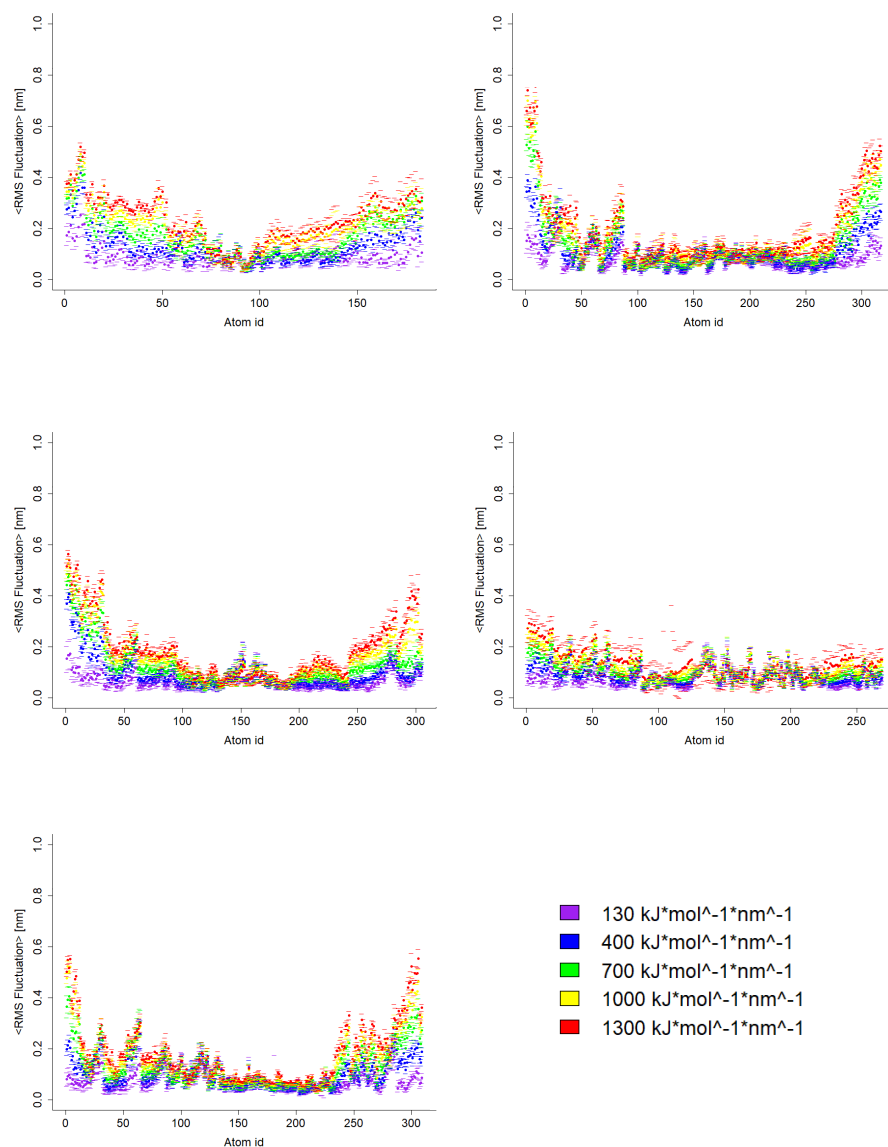


Figure 3.3.14 The change in RMS Fluctuations for each atom in a peptide. Each panel corresponds to each stretched peptide. The legend lists constant forces used to stimulate stretching. The panels will be described from left to right, from up to down. The upper left panel showcases results acquired for *Ala-only* chain. The upper right panel showcases results acquired for *Neck linker*. The middle left panel showcases results acquired for *no-Asparagine* chain. The middle right panel showcases results acquired for *no-Lysine* chain. The bottom panel showcases results acquired for *no-Proline* chain. The data has been acquired in $2 \cdot 10^3$ runs in **GROMACS** 2018. The estimation of RMS Fluctuations has been achieved by calculating values for each run and then averaging over the ensemble. Dotted lines signify standard deviations.

Chapter 4

Free energy difference estimation on the basis of Jarzynski equality and Crooks fluctuation theorem

4.1 Methods

4.1.1 Defining free energy for ensembles of conformational states

Free energy G of a thermodynamic system (*e.g.*, a molecule) for which there are N possible conformational states, can be given as

$$G = \langle U \rangle - TS, \quad (4.1)$$

where $\langle U \rangle$ is an average energy of N possible conformational states and contributes towards the growth of G , while temperature T and entropy S limit it. In an ensemble described as such, we define the average energy $\langle U \rangle$ thusly:

$$\langle U \rangle = \sum_i^N p_i U_i. \quad (4.2)$$

Here, U_i is the energy of a singular i^{th} conformational state from the ensemble, while p_i is a probability of molecule being found in the i^{th} conformational state. Energy U may be considered f.i. internal energy E in canonical ensemble or potential energy ($E + pV$; p - pressure, V - volume) in isothermal-isobaric ensemble.

If the partition function Z is

$$Z = \sum_i^N e^{-\beta U_i}, \quad (4.3)$$

then we can state that the probability p_i is

$$p_i = \frac{1}{Z} e^{-\beta U_i} \quad (4.4)$$

with $\beta = \frac{1}{k_B T}$ and k_B taken to mean Boltzman constant. Finally, entropy S of the ensemble can be phrased in terms of probabilities of conformational states:

$$S = -k_B \sum_i^N p_i \ln p_i. \quad (4.5)$$

As a result, we can rephrase free energy G as

$$G = \sum_i^N p_i U_i + k_B T \sum_i^N p_i \ln p_i. \quad (4.6)$$

A couple of routine steps of equation manipulations lead to following result:

$$G = -k_B T \ln Z = -\frac{1}{\beta} \ln Z. \quad (4.7)$$

As we can see, free energy G depends on the partition function Z in a straightforward manner. For large systems, like those that are perused in **MD** simulations, finding the exact form of the partition function is exceedingly difficult. It is possible to consider the overall partition function Z as a sum of a partition function Z_{HV} and a number of partition functions Z_j :

$$Z = Z_{HV} + \sum_j^M Z_j, \quad (4.8)$$

where $M < N$. The partition functions Z_j can be tied to the M energy wells of molecule energy U , while the rest of the potential energy's profile is subsumed into the partition function Z_{HV} . Each of M energy wells will consist of its own ensemble of conformational states. In this case, if an **MD** simulation will be constrained to only one energy well, it will sample only from a given Z_j . If a molecule is then moved to another energy well, conformational states will be sampled from an ensemble related to that one during an **MD** simulation. Still, the accuracy of sampling is difficult to ascertain and that makes it challenging to gauge G for each of ensembles. In order to compare free energy G between two ensembles of conformational states, it is possible to estimate free energy difference ΔG directly.

4.1.2 Free energy difference

The general relation between ΔG and work W performed on the system can be given as

$$\Delta G \leq W, \quad (4.9)$$

while free energy difference ΔG is itself defined as:

$$\Delta G = G_{II} - G_I, \quad (4.10)$$

where G_I and G_{II} are free energies of some ensembles I and II . If equation 4.7 is taken into consideration, the equation 4.10 becomes

$$\Delta G = -k_B T \ln \frac{Z_I}{Z_{II}}. \quad (4.11)$$

Z_I and Z_{II} are partition functions relevant to ensembles I and II .

4.1.3 Jarzynski equality

An additional difficulty with the estimation of ΔG in case of pulling experiments like **AFM** or *Optical tweezers* (and **MD** simulations that aim towards reproducing these experiments) arises from the fact, that changes are not necessarily done slowly enough to keep the process as a *quasi-static* one. This limitation can be overcome thanks to Jarzynski equality[52, 51], which is a relation that holds true regardless of the speed of the process. It is usually given as

$$\overline{e^{-\beta W}} = e^{-\beta \Delta G} \quad (4.12)$$

and sometimes as

$$\Delta G = -\frac{1}{\beta} \ln(\overline{e^{-\beta W}}). \quad (4.13)$$

If the hamiltonian $H(r, m\dot{r}, \lambda)$ of the system is parameterized by a certain quantity traditionally denoted as λ (in our case it is *end-to-end* distance of a peptide) that switches between starting value λ_0 and final value λ_s in time t_s , work W appearing in the equations 4.12 and 4.13 is defined thusly:

$$W = \int_0^{t_s} dt \frac{\partial \lambda}{\partial t} \frac{\partial H(r, m\dot{r}, \lambda)}{\partial \lambda}. \quad (4.14)$$

Position r and momentum $m\dot{r}$ variables describe the phase space of the system. Jarzynski had shown[52] that work W and subsequently ΔG , according to equation 4.13, converge to a well-known expression in the limit of t_s nearing infinity and process being *quasi-static*, as well as to another well-attested expression in the limit of immediate λ switching (t_s nearing 0 seconds)[109, 60]. More importantly, Jarzynski equality can be used to find ΔG between the conformational states for any t_s in-between these two limits.

4.1.4 Crooks fluctuation theorem

Taking into consideration the implications of Jarzynski equality, Crooks developed an expression connecting ΔG with work distributions of forward and reverse processes[21]. In his derivation, Crooks made use of the fluctuation theorem, a concept introduced in 1993[25]. The fluctuation theorem is generally presented in a following manner:

$$\frac{p(\sigma)}{p(-\sigma)} = e^{\sigma \cdot t}. \quad (4.15)$$

It evaluates a change in time t of a ratio between the probability $p(\sigma)$ of time-averaged irreversible entropy production σ and the probability $p(-\sigma)$ of time-averaged entropy production of opposite value, $-\sigma$. It can be inferred from the equation 4.15 that a chance of observing an entropy decrease is non-zero for microstates, even though the second law of thermodynamics still holds for the macrostate of a system. Assuming microscopic reversibility, Crooks restated the fluctuation theorem, relating ΔG to probability distributions of forward and reverse work of a process involving λ -switching. Firstly, the entropy production ω of a single realization of a λ -switching in time t_s was defined from canonical ensemble as

$$\omega = k \ln p_{-t_s} - k \ln p_{t_s} - \beta Q, \quad (4.16)$$

where Q is energy coming from the heat bath, $-k \ln p_{t_s}$ and $-k \ln p_{-t_s}$ are entropies of conformational states reached in forward and reverse instances of λ -switching. Definition of the entropy of a single conformational state is in agreement with equation 4.5. In case of isothermal-isobaric ensemble, additional term $-\beta p \Delta V$ must

be added. With that, the fluctuation theorem may be given in a new form, using probability distributions $P_{fwd}(\omega)$ and $P_{rev}(\omega)$ of forward and reverse entropy production:

$$\frac{P_{fwd}(\omega)}{P_{rev}(-\omega)} = e^{\omega}. \quad (4.17)$$

Now, the equation 4.4 can be used as a formula for probabilities p_{t_s} and p_{-t_s} of a particular forward and reverse trajectory during the process of λ -switching. Partition functions Z_{t_s} and Z_{-t_s} can be presented as functions of G by transforming equation 4.7. Since from the above follows that the difference of microstates' entropies

$$k \ln p_{-t_s} - k \ln p_{t_s} = \beta \Delta U - \beta \Delta G, \quad (4.18)$$

entropy production ω can be given as

$$\omega = \beta W - \beta \Delta G, \quad (4.19)$$

remembering that $\Delta E = W + Q$. Finally, we arrive at the aforementioned form of the Crooks fluctuation:

$$\frac{P_{fwd}(W_d)}{P_{rev}(-W_d)} = e^{\beta(W - \Delta G)}, \quad (4.20)$$

where $P_{fwd}(W_d)$ and $P_{rev}(-W_d)$ are probability distributions of forward and reverse dissipative work W_d of the process. The dissipative work refers to the subset of all work realizations, for which the inequality $W \geq \Delta G$ is fulfilled. The amount of work, that exceeds ΔG , feeds entropy. The W_d distributions will be equal, when $W = \Delta G$, which means that the value of W at the crossing point between the two distributions is equal to ΔG [21, 22].

4.1.5 Practical considerations and alternative free energy difference estimators

While the Jarzynski equality is well established, the precision with which it estimates ΔG depends on several factors. The ensemble of work realizations must be well averaged, at least 10^3 pathways must be gathered for square root of mean square error to fall well below $0.5 k_B T$ ("Fig. 5" in [32]). Since our ensembles contain 10^4 simulations each, we are well covered. Additional obstacle arises however, when speed of λ -switching becomes high. It has become apparent that, because of *hysteresis* effect, an estimator based purely on the Jarzynski equality overestimates the value of ΔG [59]. The growth of dissipative work leads to the growth of the Jarzynski estimator's error. This becomes evident, when comparing ΔG estimated with the Jarzynski estimator for forward and reverse processes ("FIG. 9" in [59] and "Figure 5" in [5]). The estimator based on the Crooks fluctuation theorem - the crossing point of work distributions - performs well, but only if distributions actually cross. When distributions are far away from each other, it merely provides upper and lower bound on ΔG value[59]. Two possible solutions have been considered. The first one is to make use of the fluctuation-dissipation theorem estimator ΔG_{FD} :

$$\Delta G_{FD} = \langle W \rangle - \frac{1}{2} \beta \sigma_W^2. \quad (4.21)$$

The validity of this approximation is dependant upon the work distribution being Gaussian[32] and the variance σ_W^2 of forward and reverse distributions being equal[59]. The second solution is to employ the Bennett estimator:

$$\Delta G_{Bennett} = \frac{\Delta G_{fwd} - \Delta G_{rev}}{2} \quad (4.22)$$

Here, ΔG_{fwd} and ΔG_{rev} are free energy values calculated from forward and reverse processes respectively and defined by the equation 4.13. The Bennett estimator has been shown to give good approximations regardless of the λ -switching speed[59].

In a protocol employed with our **MD** simulations, the work values were calculated following the approximation found in [20] (equation (2) therein). The **MD** simulations were performed in coupled temperature and under coupled pressure, modelling isothermal-isobaric ensemble.

4.2 Results

Since λ -switching has been defined in numerous experiments[3, 27] as manipulating *end-to-end* distance of various biopolymers, the crossing point between the probability distributions of forward and reverse dissipative work W_d has been shown to estimate well ΔG between folded and unfolded structures of biomolecules. Similar scheme has been employed in the **MD** simulations that constitute the main part of research done for this thesis. Initially, ΔG values have been estimated between stretched and relaxed structures of the *neck linker* and also between stretched and relaxed structures of the *Ala-only* chain. Subsequently, the scheme was extended to *no-Asparagine*, *no-Lysine* and *no-Proline* chains. The work and ΔG values, calculated from the resulting trajectories of these **MD** simulations, have been compared with the results from similar experiments. The aim was to ascertain whether the ensembles of the calculated work values are within sensible range of work values reported in the aforementioned experiments. While ΔG estimation should not be affected in principle (if not in practice[59]), when it comes to work values, several factors may influence these assessments - experiment's duration, the magnitude of the stretching force, the way the force is applied and the like. Nevertheless, we do get some general hints regarding expected range of work values for biopolymers operating at nanoscale. Unfolding of DNA hairpins [4], that we should expect work values of order of tens of $k_B T$ (where $k_B T$ can be given as approximately $4.14 \text{ pN} \cdot \text{nm}$). In another instance, unfolding and then refolding of protein barnase by way of optical tweezers yields work values of exceeding $1000 \text{ } k_B T$ [5]. At the other end of this range are the experiments[2] in which 18 residue-long pVEC peptide variants are pulled through the POPE bilayer. The maximum work in this case reached the range of $797 - 919 \text{ kcal/mol}$. In Fig.4.2.1 and Fig.4.2.2 we can observe that the work distributions, obtained with the highest constant stretching force used in simulations, do not cross in cases of the *neck linker* and the *no-Lysine* chain. As a result, using the crossing point of distributions for ΔG estimation is not reliable. Since in all cases the variance of forward and reverse distributions is different, the ΔG_{FD} estimator won't be precise as well. The distributions of work are Gaussian, which suggests that the peptides model spring-like behaviour well.

Peptide type	$\langle W \rangle_{stretch} [k_B T]$	$\langle W \rangle_{relax} [k_B T]$	Crooks $\Delta G [k_B T]$	$\Delta G_{Bennett} [k_B T]$
<i>neck linker</i>	1569.9 ± 24.9	1211.6 ± 68.2	not applicable	1390.75
<i>Ala-only</i>	949.7 ± 23.9	696.0 ± 81.5	882 ± 0.1	822.85
<i>no-Asparagine</i>	733.7 ± 30.5	651.0 ± 63.3	690.5 ± 0.1	692.35
<i>no-Lysine</i>	985.6 ± 40.4	388.0 ± 66.7	not applicable	686.8
<i>no-Proline</i>	1056.8 ± 26.8	888.0 ± 63.9	997.6 ± 0.1	972.4

Table 4.1: Distribution means of stretched and relaxed chain ensembles of the original *neck linker*, the *Ala-only* peptide and 3 modified (mutant-like) structures, with the stretching force value set to $1300 \text{ kJ} \cdot \text{mol}^{-1} \cdot \text{nm}^{-1}$.

In Table 4.1 we can see the comparison of the Crooks estimator and Bennett estimator and we can conclude that the results are consistent. The maximum ΔG value of $1390.75 \text{ } k_B T$ is reached by the *neck linker*, while ΔG values of other peptides do not cross the $1000 \text{ } k_B T$ threshold. It has been shown that a friction emerging

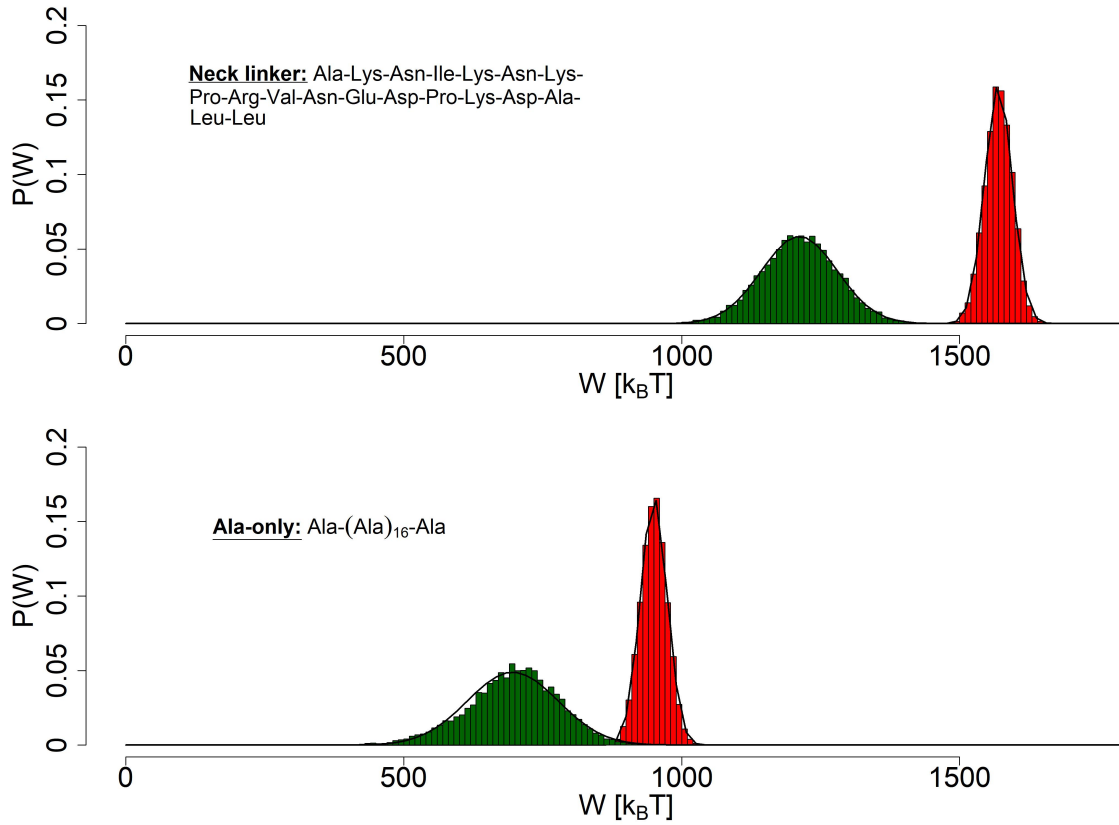


Figure 4.2.1 Reconstruction of probability density functions (PDFs) of work done on the *end-to-end* distance in pulling *in-silico* experiments of the modeled peptides (the stretching force value set at $1300 \text{ kJ} \cdot \text{mol}^{-1} \cdot \text{nm}^{-1}$). The Gaussian curves are fitted to the derived histograms of the *neck linker* (top) and *Ala-only* peptide (bottom). Crossing points of the curves are at $1462.4 k_B T$ and $882 k_B T$, respectively. The ensembles of stretched structures are red, while the ensembles of relaxed structures are green colored.

from pulling a short peptide through water affects the work values in a linear manner, its rise tied to the rise of the pulling speed[76]. It leads to the conclusion that ΔG values we have estimated here are within reasonable expectations. The results may seem to imply that the *neck linker* stretches more effectively than other peptides and returns to the relaxed conformation faster than all the other chains except for the *no-Lysine* peptide. The *no-Lysine* chain's ability to retract seems to even surpass that of the original chain. Overall, this conclusion seems to be in line with *neck linker's* documented elasticity[35]. It is well to consider, however, that each structure started from a equilibrated structure with different *end-to-end* distance. In other words, starting value of λ , λ_0 , is different in all our simulated forward work ensembles. In order to remove any bias that may arise from that, we have to take into account, we may look at the ΔG values in relation with fractional extension of the peptide. We define fractional extension as a ratio of the extension measured from the relaxed mean *end-to-end* distance to the rest of the path (L minus the relaxed mean *end-to-end* distance, where L is the contour length of the chain in question). We assume that the contour length per amino acid is equal to $4.0 \text{ \AA}$ [1]. The relation between fractional extension and ΔG is shown in Fig.4.2.3. Here, I have included ΔG values estimated for

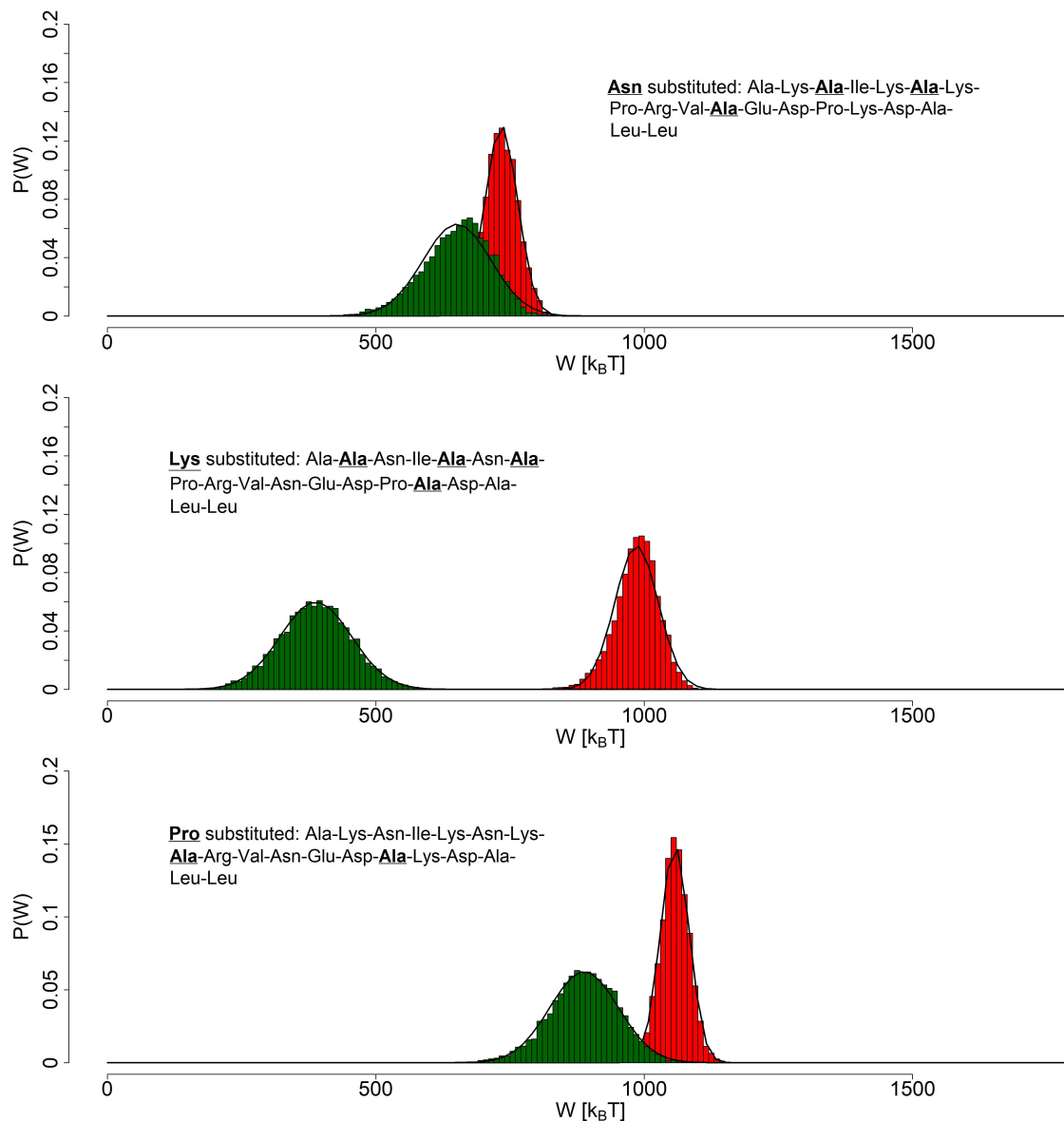


Figure 4.2.2 Reconstruction of probability density functions (PDFs) of work done on the *end-to-end* distance in pulling *in-silico* experiments on the modified linker chains (the stretching force value set at $1300 \text{ kJ} \cdot \text{mol}^{-1} \cdot \text{nm}^{-1}$). The Gaussian curves are fitted to the derived histograms of the *no-Asparagine* peptide (top), the *no-Lysine* peptide (middle) and the *no-Proline* one (bottom). The ensembles of stretched structures are red colored, while the ensembles of relaxed structures are green.

all simulated experiments described in Section 3.2 of the thesis.

Since all peptides, *neck linker* in particular, are considered to model spring-like behaviour, the quadratic

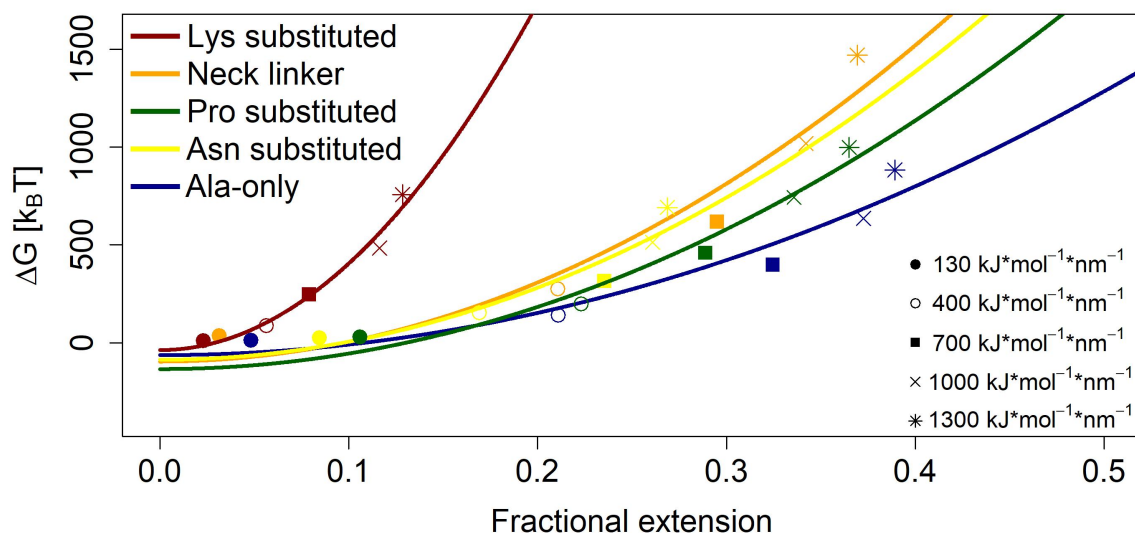


Figure 4.2.3 ΔG expressed as a function of the fractional extension for the *neck linker*, *Ala-only* peptide, *no-Asparagine* chain (*Asn substituted*), *no-Lysine* chain (*Lys substituted*) and *no-Proline* chain (*Pro substituted*). Several different ΔG values have been acquired by repeating the *in silico* experiments while using different values of a constant stretching force equal to 130, 400, 700, 1000 and 1300 $\text{kJ} \cdot \text{mol}^{-1} \cdot \text{nm}^{-1}$. Parabolic curves have been fitted to the plots. Quadratic coefficients of fitted plots are: *Lys substituted*: $44.3 \cdot 10^3 k_B T$; *neck linker*: $10.1 \cdot 10^3 k_B T$; *Asn substituted*: $9.2 \cdot 10^3 k_B T$; *Pro substituted*: $7.9 \cdot 10^3 k_B T$; *Ala-only*: $5.4 \cdot 10^3 k_B T$. They represent the effective stiffness of the stretched structure. The intercept coefficients are: -37.3, -95.8, -86.6, -134.4 and -63.1 $k_B T$, respectively.

curves have been fitted to the data points. The simulated peptides are consistent with analysis of single polymer dynamics presented elsewhere (see e.g. ref.[66]). The highest coefficient describes the curve fitted to the *no-Lysine* peptide. It is the most resistant to stretching process by far. The *neck linker* curve is characterised by the 2nd highest coefficient, while the *Ala-only* peptide curve is the lowest. It can be observed, while looking at the Fig.4.2.3 and taking note of all the quadratic coefficient values, that the curves of the *neck linker* and the *Ala-only* peptide set limits to an area. The *neck linker*, the chain present in a properly functioning biostructure like kinesin, is the upper boundary, while the *Ala-only* peptide, the most mundane chain, is the lower boundary. The most mundane chain is also the least effective spring. The *no-Asparagine* and *no-Proline* curves are placed between them. A decrease in effectiveness seems to occur when either *Proline* residues or *Asparagine* residues are removed. With that being said, if we compare the curve of the *neck linker* and the curves of mutated peptide, then the removal of *Asparagine* caused the least amount of change in quadratic coefficient. On the other hand, the *no-Lysine* curve is far to the left relative to the other curves. The staggering difference in the values of quadratic coefficients highlights this fact. While the coefficient value of the *no-Lysine* curve is close to $45 \cdot 10^3 k_B T$, the coefficient values of the remaining curves are, more or less, in the range of $5 \cdot 10^3 - 10 \cdot 10^3 k_B T$. If we continue the line of reasoning, where higher quadratic coefficient means better effectiveness as a spring, it leads towards the conclusion that the *no-Lysine* chain is the most effective spring. In other words, while other residue substitutions caused a decrease in effectiveness, the substitution of *Lysine* residues improves it. It would be tempting to speculate from these observations that the removal of *Lysine* residues may bring an improvement

to *neck linker*'s performance within the context of kinesin motor's movement along microtubules. However, the alternative where the *neck linker*'s effectiveness is attuned to the requirements of the biological system, in which *neck linker* can be found, seems more likely. The *poly-alanine* regions (the *Ala-only* chain can be classified as belonging to that category) also exist in biological context[77]. One can suspect, that there is a viable range of stiffness for springs in biomolecules and that the area of the graph in Fig. 4.2.3, that we discussed above, is part of it. It is possible that the *no-Lysine* chain is outside of that range.

Intriguingly, absence of *Lysine* in the chain has profound effect, while absence of *Asparagine* is negligible, despite the fact that in the side chains of both amino acids a single amine group is present. In the side chain of *Lysine* four methylene groups precede the amine group. The amine group is the end group of the side chain. The side chain of *Asparagine* starts with a single methylene group, a carbonyl group follows and the amine group is last. It seems that the amine group from *Lysine*'s side chain can reach further away from the backbone and potentially experience more contacts. Larger number of contacts e.g. taking part in relatively short-lived hydrogen bonds (amine groups can be both donors and acceptors) may smooth the stretching process. The *Asparagine* residues as well as the *Lysine* residues are placed closer to N-terminus on average, so the possible effect from placement on their performance shouldn't be too pronounced.

Chapter 5

Conclusions

This thesis presents results of my investigations into elastic properties of a spring-like amino acid sequence connecting two separate domains (heads) of kinesin. The focus has been placed on changes in the structure of the motor protein's fragment as a response to both a perturbation in the form of a stretching force and residue substitutions in the sequence (point mutations). In my numerical simulations all atoms of the simulated molecule have assigned coordinates and Newton's equations of motion are solved for each of them. The results presented in this thesis confirm that arbitrary changes made to the part of kinesin's sequence, labeled as *neck linker*, have a clear impact on its mechanical properties. The modified chains were created by substituting all residues of a certain type with *Alanine* for each of them in order to model point mutations. They display different structures in equilibrium, as evidenced by having a different *end-to-end* distance. Their reaction to the stretching force has been studied by analyzing their averaged trajectories - this part of the research has not been published previously elsewhere.

In order to assess the mechanical properties that were under investigation, I have designed a completely new protocol for simulating rapid stretching in **MD** runs, happening far from equilibrium, making use of a constant force, which causes varying speed of the process. Thanks to that I was able to gather several huge ensembles of non-equilibrium **MD** trajectories. As a result, I managed to track dynamic changes, over the course of the whole stretching process, in potential energy, Lennard-Jones potential, **RMSD**, radius of gyration *etc.*

The concepts arising from Jarzynski equality and Crooks Fluctuation Theorem were used to estimate free energy differences between starting and final conformational states of stretched peptides. The huge size of the ensembles was once again critical for achieving a great precision in free energy difference estimation with the Bennett estimator. The free energy differences of *neck linker*'s stretching and the 'mutated' peptides' stretching have been additionally estimated using the Bennett estimator. This result have not been published previously elsewhere. As far as I am aware, free energy difference-fractional extension profiles, that I have acquired, are the very first instance, where free energy difference values of non-equilibrium stretching experiments have been tied to a number of specific and discrete residue substitutions. In all cases there is a clear difference in the response of 'mutated' peptides to stretching forces, when compared to the original, unmodified *neck linker* structure. The absence of *Lysine* residues affects the dynamics of the chain in the most significant way. The changes in free energy difference associated with the absence of *Lysine* also happen in an unexpected opposite direction, in contrast to other mutations. The *no-Lysine* chain is also an outlier in trajectory analysis. The much higher quadratic coefficient of *no-Lysine*'s free energy difference profile curve in Fig.4.2.3 means that *no-Lysine* peptide is by far the stiffest spring among tested structures. The fact that its stiffness is so overwhelmingly higher than the stiffness of *neck linker* indicates that this peptide must be beyond reasonable range of stiffness that would allow to preserve biological functions. In other words, it would appear that the presence of *Lysine*

in the *neck linker* sequence is more crucial to the proper functioning of the linker than that of the other amino acids tested here, *Asparagine* and *Proline*.

It remains undetermined, if and how the strength of the effect of the substitutions is dependent upon a number of substitutions and their placement within the sequence. Further studies, both experimental and theoretical, would be necessary to assert, whether the effects of each substitution within a single modified sequence add up in straightforward fashion or if the combined effects are more nuanced. Searching for the possible far reaching effects of point mutations in the *neck linker* sequence on the efficiency of kinesin's stepping mechanism could also be an avenue of research to pursue.

Bibliography

- [1] Sri Rama Koti Ainavarapu, Jasna Brujić, Hector H Huang, Arun P Wiita, Hui Lu, Lewyn Li, Kirstin A Walther, Mariano Carrion-Vazquez, Hongbin Li, and Julio M Fernandez. Contour length and refolding rate of a small protein controlled by engineered disulfide bonds. *Biophysical journal*, 92(1):225–233, 2007.
- [2] Ihsan Omur Akdag and Elif Ozkirimli. The uptake mechanism of the cell-penetrating pvec peptide. *Journal of Chemistry*, 2013, 2013.
- [3] A. Alemany, M. Ribezzi-Crivellari, and F. Ritort. From free energy measurements to thermodynamic inference in nonequilibrium small systems. *New Journal of Physics*, 17(7):075009, 2015.
- [4] A. Alemany and F. Ritort. Fluctuation theorems in small systems: extending thermodynamics to the nanoscale. *Europhys. News*, 41:27–30, 2010.
- [5] Anna Alemany, Blanca Rey-Serra, Silvia Frutos, Ciro Cecconi, and Felix Ritort. Mechanical folding and unfolding of protein barnase at the single-molecule level. *Biophysical journal*, 110(1):63–74, 2016.
- [6] L.A. Amos. Kinesin from pig brain studied by electron microscopy. *J. Cell Sci.*, 87:105–111, 1987.
- [7] Ivet Bahar, Ali Rana Atilgan, and Burak Erman. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding and Design*, 2(3):173–181, 1997.
- [8] H. J.C. Berendsen, D. van der Spoel, and R. van Drunen. Gromacs: a message-passing parallel molecular dynamics implementation. *Computer Physics Communications*, 91(1-3):43–56, 1995.
- [9] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000.
- [10] Helen Berman, Kim Henrick, and Haruki Nakamura. Announcing the worldwide protein data bank. *Nature Structural Biology*, 10:980, 12 2003.
- [11] C. Bouchiat, M. D. Wang, J. Allemand, T. Strick, S. M. Block, and V. Croquette. Estimating the persistence length of a worm-like chain molecule from force-extension measurements. *Biophysical Journal*, 76:409–413, 1999.
- [12] Bernard R Brooks, Charles L Brooks III, Alexander D Mackerell Jr, Lennart Nilsson, Robert J Petrella, Benoît Roux, Youngdo Won, Georgios Archontis, Christian Bartels, Stefan Boresch, et al. Charmm: the biomolecular simulation program. *Journal of computational chemistry*, 30(10):1545–1614, 2009.

- [13] Stephen K. Burley, Helen M. Berman, Wah Chiu, Wei Dai, Justin W. Flatt, Brian P. Hudson, Jason T. Kaelber, Sagar D. Khare, Arkadiusz W. Kulczyk, Catherine L. Lawson, Grigore D. Pintilie, Andrej Sali, Brinda Vallat, John D. Westbrook, Jasmine Y. Young, and Christine Zardecki. Electron microscopy holdings of the protein data bank: the impact of the resolution revolution, new validation tools, and implications for the future. *Biophysical Reviews* 2022 14:6, 14:1281–1301, 12 2022.
- [14] Stephen K. Burley, Helen M. Berman, Jose M. Duarte, Zukang Feng, Justin W. Flatt, Brian P. Hudson, Robert Lowe, Ezra Peisach, Dennis W. Piehl, Yana Rose, Andrej Sali, Monica Sekharan, Chenghua Shao, Brinda Vallat, Maria Voigt, John D. Westbrook, Jasmine Y. Young, and Christine Zardecki. Protein data bank: A comprehensive review of 3d structure holdings and worldwide utilization by researchers, educators, and students. *Biomolecules*, 12, 10 2022.
- [15] Stephen K Burley, Helen M Berman, Gerard J Kleywegt, John L Markley, Haruki Nakamura, and Sameer Velankar. Protein data bank (pdb): the single global macromolecular structure archive. *Protein crystallography: methods and protocols*, pages 627–641, 2017.
- [16] David A Case, Thomas E Cheatham III, Tom Darden, Holger Gohlke, Ray Luo, Kenneth M Merz Jr, Alexey Onufriev, Carlos Simmerling, Bing Wang, and Robert J Woods. The amber biomolecular simulation programs. *Journal of computational chemistry*, 26(16):1668–1688, 2005.
- [17] Ryan B Case, Sarah Rice, Cynthia L Hart, Bernice Ly, and Ronald D Vale. Role of the kinesin neck linker and catalytic core in microtubule-based motility. *Current Biology*, 10(3):157–160, 2000.
- [18] Srirupa Chakraborty and Wenjun Zheng. Decrypting the structural, dynamic, and energetic basis of a monomeric kinesin interacting with a tubulin dimer in three atpase states by all-atom molecular dynamics simulation. *Biochemistry*, 54(3):859–869, 2015.
- [19] SW Chinn, K Hall, and T Vuong. Novel heterotrimeric kinesin-related protein purified from sea urchin eggs. *Nature*, 366(6452):268–270, 1993.
- [20] Delphine Collin, Felix Ritort, Christopher Jarzynski, Steven B Smith, Ignacio Tinoco Jr, and Carlos Bustamante. Verification of the crooks fluctuation theorem and recovery of rna folding free energies. *Nature*, 437(7056):231–234, 2005.
- [21] Gavin E Crooks. Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. *Physical Review E*, 60(3):2721, 1999.
- [22] Gavin E Crooks. Path-ensemble averages in systems driven far from equilibrium. *Physical review E*, 61(3):2361–2366, 2000.
- [23] Bank Protein Data. Crystallography: Protein data bank. *Nature New Biology*, 233:223–223, 10 1971.
- [24] Frédéric Douarche, Sergio Ciliberto, Artyom Petrosyan, and Ivan Rabbiosi. An experimental test of the jarzynski equality in a mechanical experiment. *Europhysics Letters*, 70(5):593, 2005.
- [25] Denis J Evans, Ezechiel Godert David Cohen, and Gary P Morriss. Probability of second law violations in shearing steady states. *Physical review letters*, 71(15):2401, 1993.
- [26] Marcus Fischer. Macromolecular room temperature crystallography. *Quarterly Reviews of Biophysics*, 54:e1, 2021.
- [27] R. F. Fox. Using nonequilibrium measurements to determine macromolecule free-energy differences. *Proceedings of the National Academy of Sciences*, 100(22):12537–12538, 2003.

- [28] James S. Fraser, Henry Van Den Bedem, Avi J. Samelson, P. Therese Lang, James M. Holton, Nathaniel Echols, and Tom Alber. Accessing protein conformational ensembles using room-temperature x-ray crystallography. *Proceedings of the National Academy of Sciences of the United States of America*, 108:16247–16252, 9 2011.
- [29] S. Genheden and U. Ryde. Will molecular dynamics simulations of proteins ever reach equilibrium? *Physical Chemistry Chemical Physics*, 14(24):8662–8677, 2012.
- [30] I. R. GIBBONS. Studies on the protein components of cilia from tetrahymena pyriformis. *Proceedings of the National Academy of Sciences of the United States of*, 50:1002–1010, 1963.
- [31] Ian R Gibbons and Arthur J Rowe. Dynein: A protein with adenosine triphosphatase activity from cilia. *Science*, 149:424–426, 1965.
- [32] Jeff Gore, Felix Ritort, and Carlos Bustamante. Bias and error in estimates of equilibrium free-energy differences from nonequilibrium measurements. *Proceedings of the National Academy of Sciences*, 100(22):12564–12569, 2003.
- [33] Laurent Guillaud, Sara Emad El-Agamy, Miki Otsuki, and Marco Terenzio. Anterograde axonal transport in neuronal homeostasis and disease. *Frontiers in Molecular Neuroscience*, 13:556175, 2020.
- [34] Mustapha Hamdi, Gaurav Sharma, Antoine Ferreira, and Constantinos Mavroidis. Characterization of protein based spring-like elastic joints for biorobotic applications. In *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.*, pages 1794–1799. IEEE, 2006.
- [35] V. Hariharan and W. O. Hancock. Insights into the mechanical properties of the kinesin neck linker domain from sequence analysis and molecular dynamics simulations. *Cellular and Molecular Bioengineering*, 2(2):177–189, 2009.
- [36] Nolan C Harris, Yang Song, and Ching-Hwa Kiang. Experimental free energy surface reconstruction from single-molecule force spectroscopy using jarzynski’s equality. *Physical review letters*, 99(6):068101, 2007.
- [37] M. Amanda Hartman and James A. Spudich. The myosin superfamily at a glance. *Journal of Cell Science*, 125:1627–1632, 4 2012.
- [38] K. Hinsen. Analysis of domain motions by approximate normal mode calculations. *Proteins*, 33:417–429, 1998.
- [39] K. Hinsen and M. J. Field. Analysis of domain motions in large proteins. *Proteins*, 34:369–382, 1999.
- [40] Nobutaka Hirokawa, Yasuko Noda, and Yasushi Okada. Kinesin and dynein superfamily proteins in organelle transport and cell division. *Current opinion in cell biology*, 10:60–73, 1998.
- [41] Nobutaka Hirokawa, Yasuko Noda, Yosuke Tanaka, and Shinsuke Niwa. Kinesin superfamily motor proteins and intracellular transport. *Nature Reviews Molecular Cell Biology* 2009 10:10, 10:682–696, 10 2009.
- [42] Nobutaka Hirokawa, Reiko Sato-Yoshitake, Naoto Kobayashi, K Kevin Pfister, George S Bloom, and Scott T Brady. Kinesin associates with anterogradely transported membranous organelles in vivo. *The Journal of cell biology*, 114(2):295–302, 1991.

- [43] Nobutaka Hirokawa, Reiko Sato-Yoshitake, Toshimichi Yoshida, and Takao Kawashima. Brain dynein (map1c) localizes on both anterogradely and retrogradely transported membranous organelles in vivo. *The Journal of cell biology*, 111(3):1027–1037, 1990.
- [44] Roger Williams Hockney, SP Goel, and JW Eastwood. Quiet high-resolution computer models of a plasma. *Journal of Computational Physics*, 14(2):148–158, 1974.
- [45] J. Howard. *Mechanics of Motor Proteins and the Cytoskeleton*. Sinauer Associates, Sunderland, MA, 2001.
- [46] G. Hummer and A. Szabo. Kinetics from nonequilibrium single-molecule pulling experiments. *Biophys. J.*, 85:5–15, 2003.
- [47] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.
- [48] Wonmuk Hwang, Matthew J Lang, and Martin Karplus. Kinesin motility is driven by subdomain dynamics. *Elife*, 6:e28948, 2017.
- [49] C. Hyeon and J. N. Onuchic. A structural perspective on the dynamics of kinesin motors. *Biophysical Journal*, 101(11):2749–2759, 2011.
- [50] Peter Höök and Richard B. Vallee. The dynein family at a glance. *Journal of Cell Science*, 119:4369–4371, 11 2006.
- [51] C. Jarzynski. Equalities and inequalities: Irreversibility and the second law of thermodynamics at the nanoscale. *Ann. Rev. Cond. Matt. Phys.*, 2(2):329–351, 2011.
- [52] Christopher Jarzynski. Nonequilibrium equality for free energy differences. *Physical Review Letters*, 78(14):2690, 1997.
- [53] F Jon Kull, Elena P Sablin, Rebecca Lau, Robert J Fletterick, and Ronald D Vale. Crystal structure of the kinesin motor domain reveals a structural similarity to myosin. *Nature*, 380(6574):550–555, 1996.
- [54] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives. Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society*, 118(45):11225–11236, 1996.
- [55] W. L. Jorgensen and J. Tirado-Rives. The opls [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *Journal of the American Chemical Society*, 110(6):1657–1666, 1988.
- [56] Hung Yi Kristal Kaan, David D Hackney, and Frank Kozielski. The structure of the kinesin-1 motor-tail complex reveals the mechanism of autoinhibition. *Science*, 333(6044):883–885, 2011.
- [57] John C Kendrew, G Bodo, Howard M Dintzis, RG Parrish, Harold Wyckoff, and David C Phillips. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, 181(4610):662–666, 1958.
- [58] Masahide Kikkawa. Big steps toward understanding dynein. *The Journal of Cell Biology*, 202:15, 7 2013.

- [59] Seongjin Kim, Yong Woon Kim, Peter Talkner, and Juyeon Yi. Comparison of free-energy estimators and their dependence on dissipated work. *Physical Review E*, 86(4):041130, 2012.
- [60] John G Kirkwood. Statistical mechanics of fluid mixtures. *The Journal of chemical physics*, 3(5):300–313, 1935.
- [61] A. B. Kolomeisky. Motor proteins and molecular motors: how to operate machines at the nanoscale. *Journal of Physics: Condensed Matter*, 25(46):463101, 2013.
- [62] A. B. Kolomeisky and H. Phillips Iii. Dynamic properties of motor proteins with two subunits. *Journal of Physics: Condensed Matter*, 17(47):S3887, 2005.
- [63] F. Kozielski, S. Sack, A. Marx, M. Thormählen, E. Schönbrunn, V. Biou, A. Thompson, E.-M. Mandelkow, and E. Mandelkow. The crystal structure of dimeric kinesin and implications for microtubule-dependent motility. *Cell*, 91(7):985–994, 1997.
- [64] J Lesitha Jeeva Kumari, R Jesu Jaya Sudan, and C Sudandiradoss. Evaluation of peptide designing strategy against subunit reassociation in mucin 1: A steered molecular dynamics approach. *PloS one*, 12(8):e0183041, 2017.
- [65] M. Kutys, J. Fricks, and W. O. Hancock. Monte Carlo analysis of neck linker extension in kinesin molecular motors. *PLoS Computational Biology*, 6(11):e1000980, 2010.
- [66] Folarin Latinwo and Charles M Schroeder. Determining elasticity from single polymer dynamics. *Soft Matter*, 10(13):2178–2187, 2014.
- [67] C. J. Lawrence, R. K. Dawe, K. R. Christie, D. W. Cleveland, S. C. Dawson, S. A. Endow, L. S.B. Goldstein, H. V. Goodson, N. Hirokawa, J. Howard, et al. A standardized kinesin nomenclature. *the Journal of Cell Biology*, 167(1):19–22, 2004.
- [68] Andrew R Leach. *Molecular modelling: principles and applications*. Pearson education, 2001.
- [69] Erik Lindahl, Berk Hess, and David Van Der Spoel. Gromacs 3.0: a package for molecular simulation and trajectory analysis. *Journal of Molecular Modeling annual*, 7(8):306–317, 2001.
- [70] Jan Liphardt, Sophie Dumont, Steven B Smith, Ignacio Tinoco Jr, and Carlos Bustamante. Equilibrium information from nonequilibrium measurements in an experimental test of Jarzynski’s equality. *Science*, 296(5574):1832–1835, 2002.
- [71] B. Lisowski, M. Świątek, M. Żabicki, and E. Gudowska-Nowak. Understanding operating principles and processivity of molecular motors. *Acta Phys. Pol. B*, 43(5):1073, 2012.
- [72] Hong-Lei Liu, Charles W Pemble IV, and Sharyn A Endow. Neck-motor interactions trigger rotation of the kinesin stalk. *Scientific reports*, 2:236, 2012.
- [73] Kun Ping Lu, Greg Finn, Tae Ho Lee, and Linda K Nicholson. Prolyl cis-trans isomerization as a molecular timer. *Nature Chemical Biology*, 3(10):619–629, 2007.
- [74] Vladimir N Maiorov and Gordon M Crippen. Size-independent comparison of protein three-dimensional structures. *Proteins: Structure, Function, and Bioinformatics*, 22(3):273–283, 1995.
- [75] Yutaka Maruyama and Ayori Mitsutake. Structural stability analysis of proteins using end-to-end distance: A 3d-rism approach. *J*, 5(1):114–125, 2022.

- [76] Robert Horst Meißner, Gang Wei, and Lucio Colombi Ciacchi. Estimation of the free energy of adsorption of a polypeptide on amorphous sio 2 from molecular dynamics simulations and force spectroscopy experiments. *Soft matter*, 11(31):6254–6265, 2015.
- [77] Pablo Mier, Carlos A Elena-Real, Juan Cortés, Pau Bernadó, and Miguel A Andrade-Navarro. The sequence context in poly-alanine regions: structure, function and conservation. *Bioinformatics*, 38(21):4851–4858, 2022.
- [78] Keir C Neuman and Attila Nagy. Single-molecule force spectroscopy: optical tweezers, magnetic tweezers and atomic force microscopy. *Nature methods*, 5(6):491–505, 2008.
- [79] R. Padinhateeri and G. I. Menon. Stretching and bending fluctuations of short DNA molecules. *Biophysical Journal*, 104(2):463–471, 2013.
- [80] S. Park, F. Khalili-Araghi, E. Tajkhorshid, and K. Schulten. Free energy calculation from steered molecular dynamics using Jarzynski’s equality. *J. Chem. Phys.*, 119(6):3559, 2003.
- [81] Bryce M Paschal, Howard S Shpetner, and Richard B Vallee. Map 1c is a microtubule-activated atpase which translocates microtubules in vitro and has dynein-like properties. *The Journal of cell biology*, 105(3):1273–1282, 1987.
- [82] Rebecca K Phillips, Logan G Peter, Susan P Gilbert, and Ivan Rayment. Family-specific kinesin structures reveal neck-linker length based on initiation of the coiled-coil. *Journal of Biological Chemistry*, 291(39):20372–20386, 2016.
- [83] S. Pronk, S. Páll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, et al. Gromacs 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, 29(7):845–854, 2013.
- [84] Di Qiu, Peter S Shenkin, Frank P Hollinger, and W Clark Still. The gb/sa continuum model for solvation. a fast analytical method for the calculation of approximate born radii. *The Journal of Physical Chemistry A*, 101(16):3005–3014, 1997.
- [85] Avin Ramaiya, Basudev Roy, Michael Bugiel, and Erik Schäffer. Kinesin rotates unidirectionally and generates torque while walking on microtubules. *Proceedings of the National Academy of Sciences*, 114(41):10894–10899, 2017.
- [86] Nathalie Reuter, Konrad Hinsén, and Jean-Jacques Lacapère. Transconformations of the serca1 ca-atpase: a normal mode study. *Biophysical journal*, 85(4):2186–2197, 2003.
- [87] Sarah Rice, Abel W Lin, Daniel Safer, Cynthia L Hart, Nariman Naber, Bridget O Carragher, Shane M Cain, Elena Pechatnikova, Elizabeth M Wilson-Kubalek, Michael Whittaker, Edward Pate, Roger Cooke, Edwin W Taylor, Ronald A Milligan, and Ronald D Vale. A structural change in the kinesin motor protein that drives motility. *Nature*, 402(6763):778–784, 1999.
- [88] Elena P Sablin, F Jon Kull, Roger Cooke, Ronald D Vale, and Robert J Fletterick. Crystal structure of the motor domain of the kinesin-related motor ncd. *Nature*, 380(6574):555–559, 1996.
- [89] Stefan Sack, Jens Müller, Alexander Marx, Manfred Thormählen, Eva-Maria Mandelkow, Scott T Brady, and Eckhard Mandelkow. X-ray structure of motor and neck domains from rat brain kinesin. *Biochemistry*, 36(51):16155–16165, 1997.

- [90] Oluranti Sadiku-Agboola and Emmanuel Rotimi Sadiku. Theoretical modeling of nanostructured formation in polymer blends. In *Nanostructured polymer blends*, pages 33–99. Elsevier, 2014.
- [91] Hitoshi Sakakibara and Kazuhiro Oiwa. Molecular organization and force-generating mechanism of dynein. *The FEBS Journal*, 278:2964–2979, 9 2011.
- [92] Guido Scarabelli and Barry J Grant. Mapping the structural and dynamical features of kinesin motor domains. *PLoS computational biology*, 9(11):e1003329, 2013.
- [93] L. J. Smith, X. Daura, and W. F. van Gunsteren. Assessing equilibration and convergence in biomolecular simulations. *Proteins: Structure, Function, and Bioinformatics*, 48(3):487–496, 2002.
- [94] Young-Hwa Song and Eckhard Mandelkow. Recombinant kinesin motor domain binds to beta-tubulin and decorates microtubules with a b surface lattice. *Proceedings of the National Academy of Sciences*, 90(5):1671–1675, 1993.
- [95] John Stone. *An Efficient Library for Parallel Ray Tracing and Animation*. Master’s thesis, Computer Science Department, University of Missouri-Rolla, April 1998.
- [96] Michał Świątek and Ewa Gudowska-Nowak. Delineating elastic properties of kinesin linker and their sensitivity to point mutations. *Scientific Reports*, 10(1):4832, 2020.
- [97] H. Teimouri, A. B. Kolomeisky, and K. Mehrabiani. Theoretical analysis of dynamic processes for interacting molecular motors. *Journal of Physics A: Mathematical and Theoretical*, 48(6):065001, 2015.
- [98] Monique M Tirion. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Physical review letters*, 77(9):1905–1908, 1996.
- [99] Michio Tomishige, Nico Stuurman, and Ronald D Vale. Single-molecule observations of neck linker conformational changes in the kinesin motor protein. *Nature structural & molecular biology*, 13(10):887–894, 2006.
- [100] Sotaro Uemura and Shin’ichi Ishiwata. Loading direction regulates the affinity of adp for kinesin. *Nature Structural & Molecular Biology*, 10(4):308, 2003.
- [101] Sotaro Uemura, Kenji Kawaguchi, Junichiro Yajima, Masaki Edamatsu, Yoko Yano Toyoshima, and Shin’ichi Ishiwata. Kinesin–microtubule binding depends on both nucleotide state and loading direction. *Proceedings of the National Academy of Sciences*, 99(9):5977–5981, 2002.
- [102] Ronald D. Vale, Thomas S. Reese, and Michael P. Sheetz. Identification of a novel force-generating protein, kinesin, involved in microtubule-based motility. *Cell*, 42:39, 1985.
- [103] Joy T. Yang, Robert A. Laymon, and Lawrence S.B. Goldstein. A three-domain structure of kinesin heavy chain revealed by dna sequence and microtubule binding analyses. *Cell*, 56:879–889, 3 1989.
- [104] A. Yildiz, M. Tomishige, A. Gennerich, and R. D. Vale. Intramolecular strain coordinates kinesin stepping behavior along microtubules. *Cell*, 134(6):1030–1041, 2008.
- [105] Ahmet Yildiz, Michio Tomishige, Ronald D Vale, and Paul R Selvin. Kinesin walks hand-over-hand. *Science*, 303(5658):676–678, 2004.
- [106] Z. Zhang, Y. Shi, and H. Liu. Molecular dynamics simulations of peptides and proteins with amplified collective motions. *Biophysical Journal*, 84(6):3583–3593, 2003.

- [107] Z. Zhang and D. Thirumalai. Dissecting the kinematics of the kinesin step. *Structure*, 20(4):628–640, 2012.
- [108] Wenjun Zheng and Sebastian Doniach. A comparative study of motor-protein motions by using a simple elastic-network model. *Proceedings of the National Academy of Sciences*, 100(23):13253–13258, 2003.
- [109] Robert W Zwanzig. High-temperature equation of state by a perturbation method. i. nonpolar gases. *The Journal of Chemical Physics*, 22(8):1420–1426, 1954.