Local charge to global structure in the face of disorder.

Carlos G. Oliver

Master of Science

Department of Biology

McGill University

Montreal,Quebec

June 23, 2016

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science

DEDICATION

This document is dedicated to the graduate students of the McGill University.

ACKNOWLEDGEMENTS

Acknowledgments, if included, must be written in complete sentences. Do not use direct address. For example, instead of Thanks, Mom and Dad!, you should say I thank my parents.

ABSTRACT

Abstract in English and French are required. The text of the abstract in English begins here.

ABRÉGÉ

The text of the abstract in French begins here.

TABLE OF CONTENTS

LIST OF TABLES

Table page

LIST OF ABBREVIATIONS

CHAPTER 1
Introduction

Everything existing in the universe
is the fruit of chance and necessity.

Democritus

Biological systems have evolved over billions of years to ensure their proper functioning in the face of ever changing environments. As a result, evolution has produced an immense diversity of highly robust systems/organisms of ever increasing complexity. Looking closely at these systems we find that at their heart is a complex network of molecular machines. Molecular machines are assemblies of protein macromolecules that interact in a highly coordinated manner to accomplish the various fundamental tasks to ensure survival of the organism in the eternal battle against entropy. These tasks encompass all the essential processes necessary for proper cellular functioning such as cell division, growth, nutrient transport, energy generation, immune response, etc. At the core of each of those processes is a highly complex and precise machine composed of countless interacting and inter-dependent proteins that work in concert to achieve a singular goal in a robust and flexible manner. For example, the vital process of DNA replication is executed by a large multitude of proteins that each contribute to the process of copying the genome in a timely and high fidelity manner. Complex combinatoric signalling networks ensure

maybe put DNA example in next paragraph

that DNA replication in a processive manner and under the right conditions (cell cycle timing, nutrient availability, etc.), while other components perform physical work by unwinding the DNA strand to copy, and others communicate with the replication machinery to correct any copying errors and avoid harmful mutations. All of this requires participating proteins to be able to interact with many different partners and mediate many different processes. At the core of all these interactions are protein structural elements, that through various physical mechanisms allow the cell to control the multitude of interactions that define molecular machines. The broad aim of this thesis is to improve our understanding of the physical mechanisms underlying the functional complexity of molecular machines.

Looking closer at molecular machines, we see that their primary working unit is the protein. All communication between parts and functional output of the molecular machine occurs at the protein level. This activity is driven by the structural and dynamic properties encoded in each protein's amino acid sequence. Specific spatial arrangements of peptide chains, also known as protein structure, allows for specific interactions between proteins to assemble molecular machines, recruit necessary factors and mediate necessary chemical reactions. Since the 1950s when the first X-ray crystallography protein structure was solved, we have learned a great deal about how 3D architecture and motions of the chains give rise to protein function. By capturing various conformations of folded protein domains, we have been able to infer that motions between structural conformations is the main element of control in protein function. For example, . **Fig. ??** It is important to note that X-ray crystallography sill offers only static pictures of protein structure, and provides information mostly

include figure o
yeast γ-tubulin
structure to
illustrate stable
vs IDP

example of
basic protein
structure-
rearrangement

the spatial arrangement of relatively large stable domains in proteins. Yet, as we saw with DNA replication, molecular processes are incredibly complex, and a single protein often has to play many roles, interact with various different partners and be able to do so in a rapid and controllable manner. It is therefore unlikely that such large scale and consequently slow structural motions can account for all of the precise and rapid control we observe in biological systems. A static description of proteins is not sufficient to explain the degree of functional flexibility and control that we observe. How can the same protein fulfill multiple functions and engage in many different interactions? How can molecular machines offer such precise control of functionality while counting only on a static architectures?

1.1 Disorder in proteins

One potential source of plasticity can be found when looking more closely at proteins we find that the majority contain significant levels of structural flexibility, also termed intrinsically disordered regions, or proteins (IDPs). . IDPs are segments fix this sentenc of protein, or entire proteins, that do not natively adopt any stable conformation or fold but are often functionally active and thus do not follow the classical structure-function paradigm. Instead, IDPs are highly dynamic by nature and are able to rapidly sample a wide range conformations in an almost stochastic manner. This flexibility offers the protein access to a vast pool of possible conformations with which to fine-tune and diversify its function. It can then be said that IDP function lies in the 'absence'of structure. The lack of structure in IDPs can be explained by the characteristically low sequence complexities, an enrichment for polar and charged residues over large hydrophobic amino acids which tend to favour rigid folding.

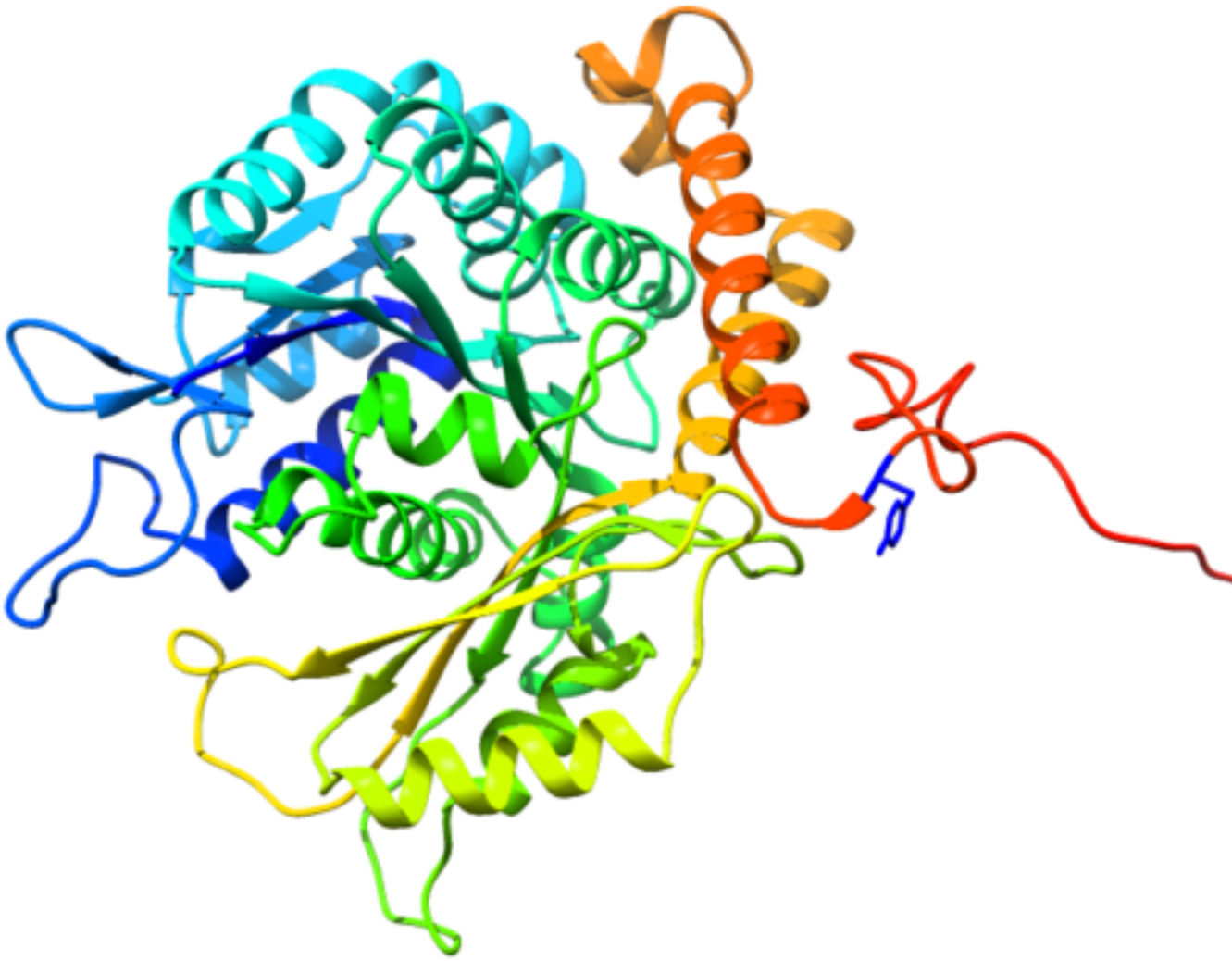


Figure 1–1: Homology model

1.1.1 Physical mechanisms of IDP function in cellular machines

Given that disorder in proteins is so prevalent it is not surprising that IDPs have been implicated in a multitude of cellular processes through NMR and mutational studies. These processes include signalling, cell cycle control, transcription, translation, ribosome assembly, chromatin organization, microtubule assembly/disassembly,

etc. Mutations in IDPs have been shown to be involved in several disease phenotypes. Interestingly, it has been shown that viral proteins use IDPs in their proteins to hijack cellular proteins and use the flexibility of IDPs to mimic host proteins and recruit host cellular machinery in order to propagate.[1] An interesting hypothesis that came from these observations is that viral proteins use IDPs to make efficient use of their smaller genomes and obtain a greater range of function from the limited number of proteins in their genomes. It is now clear that IDPs, through their lack of structure, are an important adaptive feature allowing the functional complexity and robustness we observe in molecular machines.

In this section we will give a brief account of some of the physical mechanisms of IDP function that have been described in the literature.

Phosphorylation

A key aspect of dynamic control is the ability to modulate function in a precise and reversible manner. The cell needs to be able to induce and inhibit interactions in a time and space dependent manner. To solve this problem, the cell harnesses the structural malleability of IDPs by coupling it with post translational modifications, most commonly, phosphorylation. Phosphorylation is the reversible addition of a phosphate group, which carries a negative charge, to one a tyrosine or serine amino acid by a kinase enzyme. The reverse reaction is catalyzed by enzymes called proteases which remove the phosphate group. The addition of a phosphate group introduces the potential for hydrogen bonding with itself and with other targets, and

alters the electrostatic environment of the IDP. This change can therefore bias the stochastic conformational sampling of the IDP in a particular direction, and because it is reversible, it acts as a structural switch which can then be used to modulate a large range of interactions. Not surprisingly, it has been seen in many studies that IDPs are prime targets for phosphorylation *in vivo*.

Disorder-Order transitions

The best explored mechanism of IDP function is the fold-on binding paradigm. In this case, IDPs in the free form are unstructured, and when they encounter their binding target, they undergo a folding transition (disorder to order) to form a stable complex. The lack of structure in the unbound state allows the the IDP to recognize multiple targets, and it allows the binding to be inducible instead of constitutive. A well studied example of this kind of mechanism is the binding of the transcriptional activator protein CREB and its co-activator CPB. An IDR in CREB known as KID mediates binding to CPB where upon binding, the IDP folds into a pair of helices. However, this binding process is not favoured spontaneously due to loss of entropy induced by folding. However, when the KID is phosphorylated, the phosphyl group interacts with CPB by forming hydrogen bonds which result in a negative enthalpic change that compensates for the loss of entropy and thus makes the folding reaction favourable. Because of the inducible nature of this interaction, CPB is able to also interact with other co-factors, which has been reported in the literature. [7] This is an example of how even though the association state of the IDP is ordered, without

the intrinsic disorder of the unbound state, the high entropy of IDPs acts as a barrier for binding which can be overcome in a controllable manner.

IDPs accomplish these functions via a number of physical mechanisms, many of which remain unknown and many new ones are being described.

There are many ways in which the cell harnesses the flexibility in IDPs to coordinate processes. Very often, IDPs are targets for reversible post-translational modifications such as phosphorylation. By introducing these modifications, the IDP's conformational sampling is

Fuzzy complexes

Reduce entropic cost of binding.

High specificity, low affinity –¿ high kinetic rates –¿ rapid switch-like control.

Talk about figure from Parker, 1999 entropy/enthalpy

1.2 IDP function in γ-Tubulin

1.3 Experimental question

1.4 Approach

CHAPTER 2
Theory

2.1 Molecular Dynamics Simulations

2.1.1 Fixed-temperature Molecular Dynamics

2.1.2 Replica Exchange Molecular Dynamics

2.2 Evolutionary Algorithms

CHAPTER 3
Conformational analysis of the γ-Tubulin carboxyl terminus

In this chapter we discuss the impact of local charge on the global dynamics of the γ-Tubulin C-terminus (γ-CT). In order to measure how the addition of local negative charge in an acidic polypeptide affects the conformational sampling of the γ-CT, we simulated the dynamics of various isoforms of the γ-CT using MD. By comparing results from our simulations to experimental measurements performed with NMR spectroscopy on the γ-CT, we were able to propose that local changes in charge at specific residues in the polypeptide have the ability to bias the conformational sampling of the γ-CT in such a way that may be regulating the availability of binding surfaces on γ-Tubulin.

3.1 NMR

Talk about NMR stuff here. ‘

3.2 Setting up the MD runs

Molecular Dynamics simulations (MDS) on WT and Y11D γ-CTs were carried out using MPI-enabled GROMACS 4.6.6 software[3] and a CentOS 5 high performance computational cluster. Calculations were distributed over 64 Dual Sandy Bridge 8-core, 2.6 GHz computing nodes and run under periodic boundary conditions with the OPLS-AA (Optimized Potential for Liquid Simulations ? All Atom) force field [5]. The starting γ-CT polypeptide configurations were obtained from secondary and tertiary structure predictions by RaptorX [4] and solvated using the

SPCE (extended single point charge) water model in a dodecahedral box while enforcing a minimum distance between the edge of the box and solute of 1 nanometer. The total charge of the system was neutralized by adding sodium ions to the solution. Energy minimization was carried out using a steepest descent algorithm for a maximum of 50,000 steps until a maximum force of 100 kJ/mol between atoms was achieved. A 1 nm cut-off was used for non-bonded interactions, and long-range electrostatics were calculated using a Particle Mesh Edwald Sum algorithm. The systems equilibrated under the constant NVT and NPT ensembles (288K and 1 atm) for 5 ns before the production 2 μs simulations. Post-processing of all trajectories was done using the `trjconv` module of GROMACS. Theoretical random-coil structural ensembles (10,000 conformers) were calculated based on the γ-CT primary amino acid sequence using Flexible Meccano software [6]. Translational diffusion coefficients were calculated for each structure using hydroNMR software [2]. MD conformations were grouped into percentile classes based on radius of gyration (Rg) computed using the GROMACS `g_gyrate` module. Each Rg percentile group was represented by the three structures with lowest root-mean-squared-difference RMSD values to all other structures, calculated using the GROMACS `g_rms` module. Atomic distance matrices were calculated using the GROMACS `g_mdmat` module.

3.3 Conformational sampling of γ-CT isoforms

Our NMR experiments provide evidence for a major alteration in the global dynamics of the γ-CT in the presence of the Y11D mutation characterized by collective motions involving the entire polypeptide chain occurring on the microsecond timescale. We then asked whether this phenomenon can be reproduced *in silico* using

MDS. If we are able to reproduce the transition *in silico*, the resulting MD data can be used to obtain additional insight into the structural characteristics of the conformational sampling of the γ-CT at an atomic resolution. We computed trajectories for the WT and Y11D γ-CT polypeptides by running independent 2μs GROMACS simulations, and computed the translational diffusion coefficient from the resulting MDS structural trajectories at 1 ns time steps **Fig. 3.3**. We found that the diffusion coefficient (Dc) of the WT γ-CT remains relatively constant over the total simulation time (Dc= $1.237 \times 10^{-6} \pm 1.5816 \times 10^{-8}\,\mathrm{cm^2\,s^{-1}}$ and agrees well with the NMR-derived value (Dc=$1.25 \times 10^{-6} \pm 1 \times 10^{-8}\,\mathrm{cm^2\,s^{-1}}$. Similarly to what was seen by NMR, we find that the mean $Dc$of the Y11D γ-CT polypeptide is slightly lower than that of the WT γ-CT (Dc= $1.224 \times 10^{-6} \pm 3.503 \times 10^{-8}\,\mathrm{cm^2\,s^{-1}}$). These results confirm that the γ-CT, while disordered, is more compact than a fully denatured polypeptide chain. Interestingly, between 762 to 1255 ns in the MDS, the Y11D γ-CT underwent transient excursions to less compact con-formations with significantly lower diffusion coefficients (mean Dc= $1.152 \times 10^{-6} \pm 2.0325 \times 10^{-8}\,\mathrm{cm^2\,s^{-1}}$). This sub-population is more ex-tended (i.e. diffuses more slowly) than any conformation sampled by the WT γ-CT throughout the entire MDS. While the Y11D γ-CT extended states do not overlap with the conformational ensemble of the WT γ-CT polypeptides, they do, however, lie within the conformational space expected for a typical random-coil poly-peptide, as modeled by an ensemble of 10,000 disordered extended conformers (Fig. S8) derived from the γ-CT primary sequence using the `Flexible Meccano` tool.

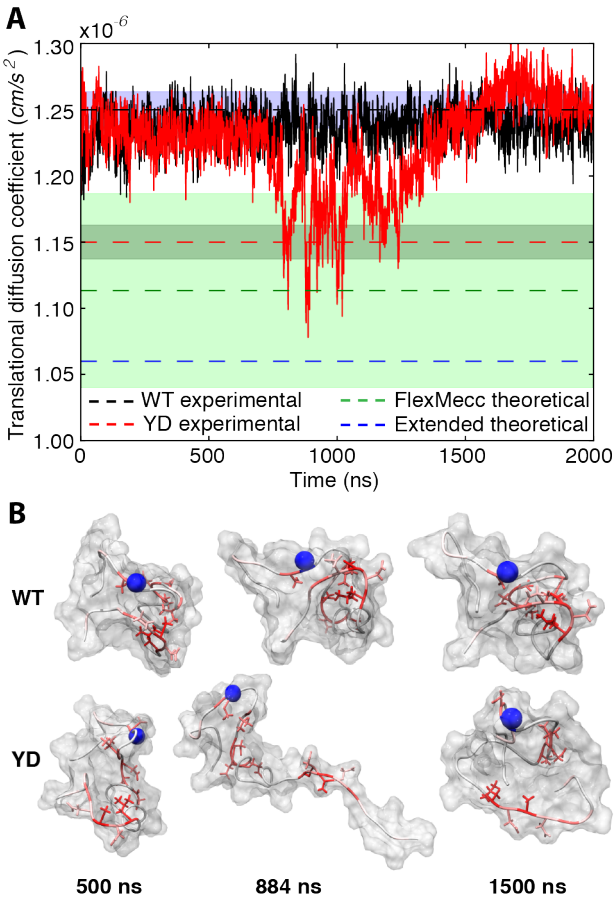


Figure 3–1: Diffusion Coefficient

In order to obtain an initial visualization of the transient opening motions experienced by the Y11D polypeptide, we extracted struc-tural models from the simulation at the time step with the lowest value of Ds for the Y11D γ-CT polypeptide (844 ns; Dc= $1.078 \times 10^{-6}\,cm^2\,s^{-1}$) as well as the time steps at 500 ns (Dc=$1.268 \times 10^{-6}\,cm^2\,s^{-1}$ for WT and $1.235 \times 10^{-6}\,cm^2\,s^{-1}$ for Y11D) and 1500 ns (Dc=$1.233 \times 10^{-6}\,cm^2\,s^{-1}$ for WT and $1.241 \times 10^{-6}\,cm^2\,s^{-1}$s for Y11D). The expansion experienced by the Y11D polypeptide is clearly observed in the 844 ns structure. We hypothesize that the global microsecond dynamics characterized by NMR for the Y11D γ-CT correspond to the transient opening motions seen by MDS, with the major state corresponding to the compact form and the minor state corresponding to the expanded form. Residues with large-magnitude dispersion profiles (i.e. those in Figure 6 with large ΔR2) have chemical shifts that are quite different in the major and the minor states. If the NMR dynamics correspond to transient expansion, these residues should experience different local environments in the collapsed and expanded states. Residues with large ΔR2 values (coloured red in Figure 7B) indeed are located in a compact cluster at 500 ns for the Y11D γ-CT and at all time steps for the WT γ-CT. However during expansion, this cluster dissociates and the residues become more solvent-exposed. This provides a possible explanation for how residues throughout a disordered polypeptide can experience a concerted, two-state, dynamical process presence of the Y11D mutation, and suggests that it is the separation of a cluster of residues located in N and C termini of the γ-CT polypeptide that drives a transition to extended conformations with a concomitant reduction of the translational diffusion coefficient.

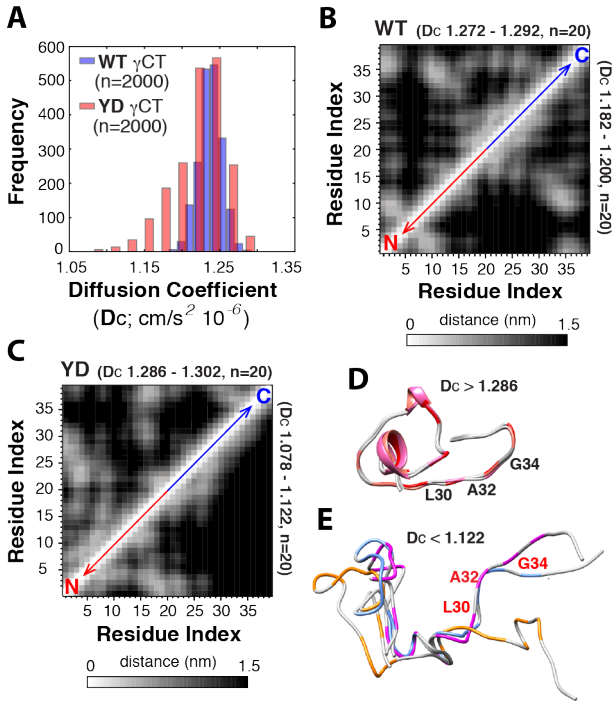
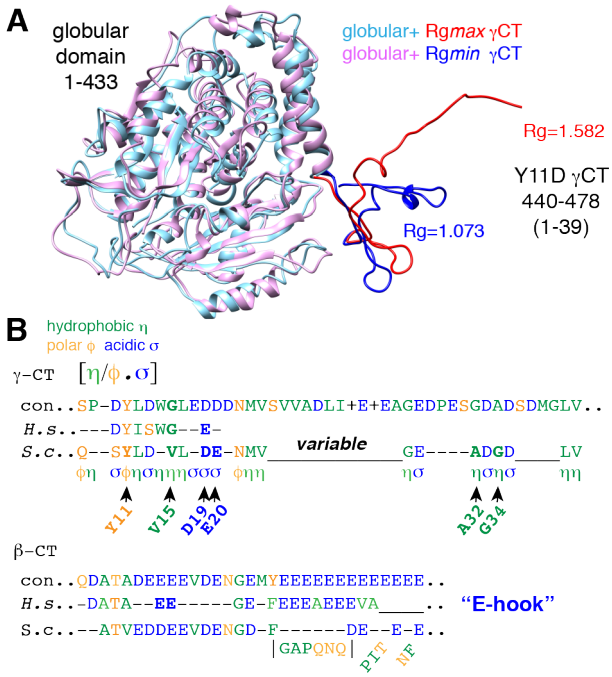


Figure 3–2: Contact Maps

To further characterize the transition observed in the Y11D γ-CT polypeptide, we chose a subset of 2000 time steps (every 1 ns) for additional analysis. We calculated theoretical diffu-sion coefficients for each structure, yielding the frequency histo-grams shown Figure 8A. The Y11D γ-CT distribution is clearly skewed compared to that of the WT with many structures exhibit-ing far slower self-diffusion and more extended conformations. The conformations within the top and bottom 1% (20 structures each) of the WT γ-CT and Y11D γ-CT Ds distributions best represent the collapsed (top 1%) and extended (bottom 1%) conformations for γ-CT polypeptides. In the case of the WT, we do not expect the upper and lower Ds subsets to substantially differ, as the WT γ-CT conformations exhibit fairly homogenous compactness overall. For Y11D γ-CT, we expect the upper Ds subset to resemble that of the WT γ-CT, while the lower Ds subset is expected to reflect the tran-sient opening process. We plotted the mean distance between alpha carbons of all pairs of residues for as contact maps for the set of collapsed (upper) and extended (lower) conformations of the WT γ-CT polypeptide (Fig. 8B) and the Y11D γ-CT polypeptide (Fig. 8C). As expected, the upper and lower Ds subsets of the WT γ-CT and the upper Ds subset of the Y11D γ-CT polypeptides show similar patterns of pair-wise contacts. In contrast, the C-terminal residues in the lower Ds subset of the Y11D γ-CT lose the majority of contacts with N terminal residues, as a consequence of the conformational expansion. Next, we isolated the three confor-mations from the upper and lower Ds subsets of Y11D γ-CT poly-peptides with the lowest all-to-all RMS, also known as centroid structures, shown in Fig. 8D,E. with large relaxation

dispersion magnitudes indicated in red. This analysis shows that the extended conformations consist of a compact N-terminus with residues lo-cated in the C-terminal region of the γ-CT, (including dynamical-ly-broadened residues L30, A32 and G34) isolated from the N-terminus and solvent-accessible. Through MDS we are able to re-produce the anomalously rapid diffusion (i.e. high compactness) of the WT and Y11D ground-state γ-CT polypeptides. Moreover, we saw that the Y¿D substitution caused relatively slow collective motions of the entire polypeptide chain, as observed by NMR.

Representative structures obtained identifying structures with lowest all-to-all RMSD values for the YD high diffusion coefficient group (D) and YD low diffusion coefficient group (E), residues with ΔR2 val-ues greater than 5 s-1 are labeled in red. Our experimental analysis of the structural properties of the γ-CT using NMR and corresponding MDS are based on the properties of the WT and Y11D γ-CT polypeptides in isolation. In order to determine whether the conformations and dynamics we observed for the isolated ??CTs are physically consistent within the context of the full-length γ-tubulin protein, we docked the min-imum Ds γ-CT model (Fig. S9) onto the globular domain of an S.c. ?-tubulin homology model and used this as an initial structure for whole protein simulations on γ?tubulin. Due to the increase in system size, simulation times were reduced to 200ns. As with the Y11D γ-CT polypeptide, the γ-CT in the whole protein simula-tion underwent exchange between extended and compact confor-mations (Fig. S10), suggesting both states are accessible in the presence of the globular domain. We found no contacts between residues in the globular domain with the 39 residues of the γ-CT throughout the 200



ns simulation (minimal distance between any pair of residues is ¿ 0.7 nm). Structures for the full protein with the γ-CT at minimum radius of gyration (1.073 nm; model S11) and maximum radius of gyration (1.582 nm; model S12) are shown in Fig. 9A.

The CTs of α- β- and γ-tubulins are enriched in acidic residues (Asp, Glu). γ-CT s across eukaryotes additionally contain clusters of hydrophobic or polar residues which are not found in α- or β-CTs. Interestingly, the residues most broadened in Y11D NMR spectra, i.e. those most affected by the compact-to-extended transition (V15, D19, E20, A32, G34), are all found in positions conserved either on a sequence level or on a physical property level (polarity/charge) in a consensus γ-CT sequence(Fig. 9B). This suggests that clusters of hydrophobic residues, including those that contribute to transitions between compact and extended confor-mations

in the S.c. D11 γ-CT, are a feature of an otherwise diverse set of γ-CT s across many eukaryotic organisms.

CHAPTER 4
MEDIEVAL

?A totally blind process can by definition lead to anything; it can even lead to vision itself.? Jacques Monod

CHAPTER 5
Conclusions

Appendix A

Here is the text of an Appendix.

Appendix B

Here is the text of a second, additional Appendix

References

[1] Norman E Davey, Gilles Travé, and Toby J Gibson. How viruses hijack cell regulation. *Trends in biochemical sciences*, 36(3):159–169, 2011.

[2] J Garcıa De la Torre, ML Huertas, and B Carrasco. Hydronmr: prediction of nmr relaxation of globular proteins from atomic-level structures and hydrodynamic calculations. *Journal of Magnetic Resonance*, 147(1):138–146, 2000.

[3] Berk Hess, Carsten Kutzner, David Van Der Spoel, and Erik Lindahl. Gromacs 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of chemical theory and computation*, 4(3):435–447, 2008.

[4] Morten Källberg, Haipeng Wang, Sheng Wang, Jian Peng, Zhiyong Wang, Hui Lu, and Jinbo Xu. Template-based protein structure modeling using the raptorx web server. *Nature protocols*, 7(8):1511–1522, 2012.

[5] George A Kaminski, Richard A Friesner, Julian Tirado-Rives, and William L Jorgensen. Evaluation and reparametrization of the opls-aa force field for proteins via comparison with accurate quantum chemical calculations on peptides. *The Journal of Physical Chemistry B*, 105(28):6474–6487, 2001.

[6] Valéry Ozenne, Frédéric Bauer, Loïc Salmon, Jie-rong Huang, Malene Ringkjøbing Jensen, Stéphane Segard, Pau Bernadó, Céline Charavay, and Martin Blackledge. Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. *Bioinformatics*, 28(11):1463–1470, 2012.

[7] Ishwar Radhakrishnan, Gabriela C Pérez-Alvarado, David Parker, H Jane Dyson, Marc R Montminy, and Peter E Wright. Solution structure of the kix domain of cbp bound to the transactivation domain of creb: a model for activator: coactivator interactions. *Cell*, 91(6):741–752, 1997.

KEY TO ABBREVIATIONS