Local charge to global structure in the face of disorder.

Carlos G. Oliver

Master of Science

Department of Biology

McGill University

Montreal,Quebec

July 7, 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

Molecular machines are assemblies of proteins that interact in a coordinated manner to solve a biological problem. These biological problems, such as coordinating chromosome separation during cell division, , etc. cover all processes essential to life. Given the necessity for survival in the face of ever changing environments, evolution has produced a large diversity of intricate molecular mechanisms for solving these problems in a flexible and robust manner. A machine that functions properly only under unchanging conditions and inputs is not likely to survive in a biological context. Therefore, at the core of each of those processes are highly complex networks of protein interactions that are able to assemble, communicate, coordinate, self-regulate and self-correct in order to accomplish the necessary task reliably. 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. The replication machinery must read environmental queues for initiating replication at the correct time, while complex combinatoric signalling networks ensure that DNA replication unfolds in a processive manner, enzymatic components perform physical work by unwinding the DNA strand for copying, and others communicate with the DNA repair machinery

get more examples

maybe put DNA example in next paragraph

to correct copying errors and avoid harmful mutations. It is clear that solving this biological problem requires the ability of participating proteins to interact with many different partners and mediate many different processes.

The way that proteins are able to control interactions and drive function is through the dynamic properties of their structural domains 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 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?

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

example of basic protein structure-rearrangement

How can molecular machines offer such precise control of functionality while counting only on a static architectures? The broad aim of this thesis is to improve our understanding of the physical mechanisms underlying the functional complexity of molecular machines.

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. [1] 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.

For a long time, these types of protein element were overlooked in favour of studying stable folding patterns. In the case of IDRs, IDRs were thought to act

[1] We will use the term IDP loosely to denote any unstructured protein element. Some works make the distinction between IDP and IDR where an IDR is an intrinsically disordered region and is not a full protein. There will be instances in the text where we will use IDR in this way.

simply as flexible linkers between structural domains without any function or their own. This was also due in part to the fact that the main tool being used for structural biology, X-ray crystallography, fails to detect patterns in unfolded chains, making it difficult to study highly dynamic elements in protein such as IDPs. Techniques that do produce information on dynamics such as NMR only developed for proteins until much later; the first protein structure solved by X-ray crystallography was in 1958 and it was not until 1984 when the first protein structure was solved using NMR. Likewise, computational tools to study protein dynamics *ab initio*, such as Molecular Dynamics (MD) were greatly limited by shortcomings in processor power. However, with large advances in experimental and computational techniques in recent years, we have been able to study the dynamic properties of IDEs in great detail, and have found that they play key roles in the functioning of molecular machines.

1.2 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.[2] 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

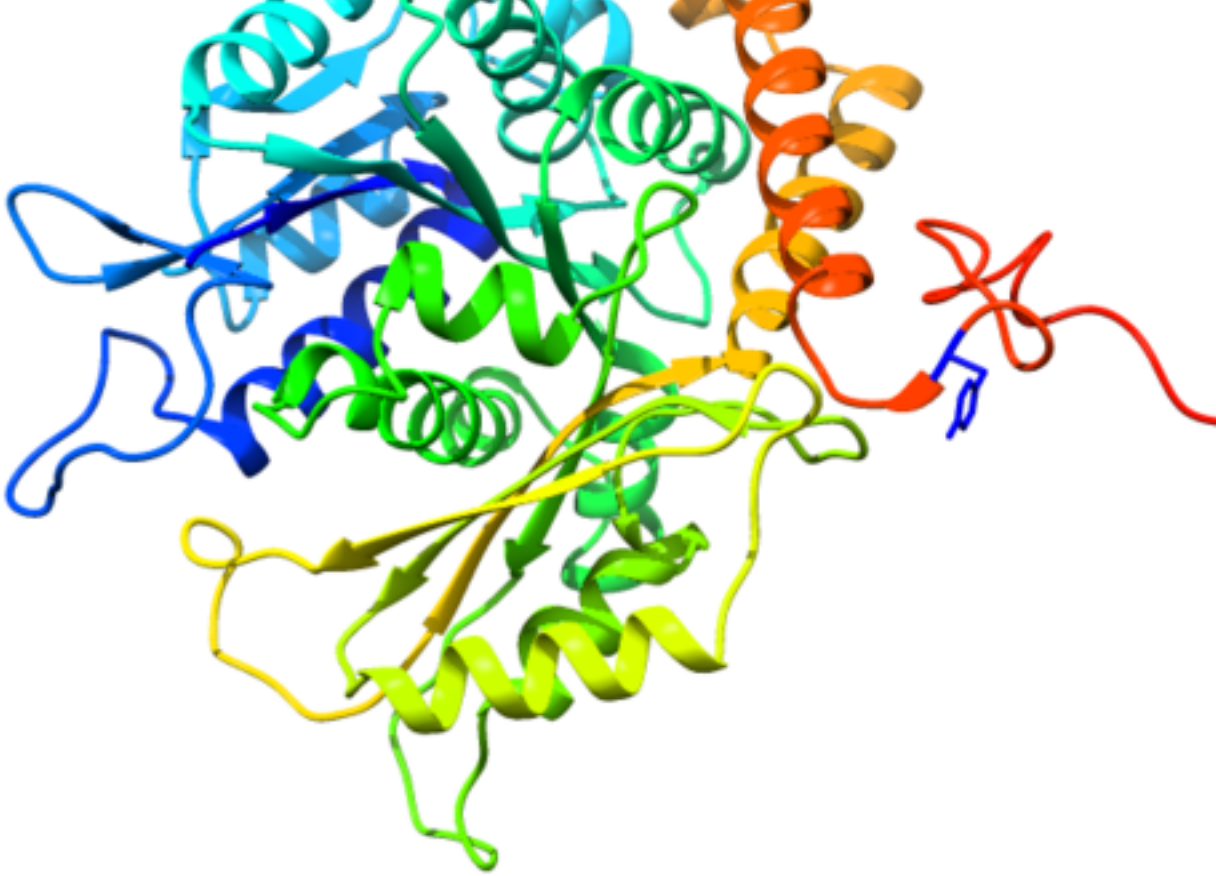


Figure 1–1: Homology model

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*. The use of phosphorylation as an information carrier has been described in various cellular systems. For example, in cell cycle control, a specific temporal sequence of interactions has to be enforced. Therefore, the sequential phosphorylation of a single target can modify the affinity for the same target to the subsequent target in the pathway. Phosphorylation can also be used to enforce thresholds, where to avoid the negative impact of interactions by coincidence, certain interactions will only be activated once an IDP has achieved a certain number of phosphorylations. ...

insert examples of phosphorylation in combinatorial, information carrying, etc.

Disorder-Order transitions

The best explored physical 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. [2] 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. [10] This is an example of how even though the association state of the

[2] If we consider the expression for Gibbs free energy, $\Delta G = \Delta H - T\Delta S$, where ΔH is the change in enthalpy, or heat energy available to the system, and ΔS is the entropy, or degree of disorder/conformational freedom available to the system. If we define our system as a protein chain, we see that a decrease in entropy which can be brought upon by the ordering of a structure will result in a positive ΔG thus making the reaction energetically unfavourable. This loss of entropy, or conformational freedom, can be overcome by a release of heat, or a sufficiently negative ΔH component which can tip the energetic balance toward $\Delta G < 0$. One might say that the decrease in entropy in the protein would be in violation of the second law of thermodynamics which states that all systems tend toward an increase in entropy. However, a reaction with a negative ΔH, in other words, a reaction releasing heat will cause an increase in entropy to its surrounding environment. In the case of protein, the process of folding will reduce generally the entropy of the protein, but if the folding has a negative ΔH it will release heat and thus increase the entropy of the water molecules surrounding it.

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 to promote interactions in a controllable manner.

'Fuzzy'interactions

Disorder to order transitions are not necessary for IDPs to confer functionality. There is a growing number of examples where IDPs are involved in functional interactions while remaining in a disordered state [**?**]. Such interactions have been labeled with the term 'fuzzy'as they maintain a heterogenous conformational ensemble throughout their lifetimes. There exist various physical mechanisms by which fuzziness, or disorder, in binding interactions confers advantages to protein function. For example, binding interactions between an IDP and a target protein where the IDP is able to form alternate contacts with its binding target can help reduce the entropic cost of binding, as well as control the accessibility of different sites on the protein for modulating interactions with different targets. [5, 4] IDPs can also play a role in interactions without making direct contacts with the binding partner. By acting as flexible linkers for folded domains [1], or as 'antennae'for [11] recruiting further interactions and stabilizing the binding of folded domains through long range interactions [13, 12]

paragraph feels too compact

It is becoming increasingly evident that nature is able to harness disorder as an adaptive mechanism for control in protein-function. Flexibility in conformational state allows cellular processes to greatly expand the functional repertoire of proteins by allowing for rapid switch-like control over interactions, expanding the functional repertoire of proteins, and fine-tuning the kinetics of interactions. It is due to all of these various physical mechanisms that the cell is able to carry out its complex tasks with such robustness and precision. Advances in this field have caused us to

reconsider the paradigm of 'one structure, one function'that has prevailed in structural biology for decades. However, this remains a relatively novel area of structural biology, and there still remain many unsolved physical mechanisms in IDPs.

1.3 IDP function in the mitotic spindle

In this section we will address the role of IDPs in controlling the function of the mitotic spindle. The mitotic spindle is a complex molecular machine composed of microtubules, force generators, effector proteins, chromosomes, and numerous effector molecules which act in a coordinated manner to accomplish the process of chromosome segregation during cell division. This process ensures that genetic material is transferred from the mother to the daughter cell in a timely manner and without errors which would in most cases result in lethality. Because cell division is a fundamental task in every cell's lifetime, the mitotic spindle has shares common design features throughout eukaryotes. The task of properly arranging and segregating chromosomes is effected ultimately by filament-like polymers known as microtubules which attach to chromosomes and physically pull the DNA to the daughter cell. In order to study the underlying mechanisms at play, we work with the mitotic spindle of the budding yeast *Saccharomyces cerevisiae* due to its minimal, yet highly conserved construction.

1.3.1 Microtubules

Microtubules are constructed of alternating pairs, or dimers, of the globular proteins α and β Tubulin. A cylindrical arrangement of tubulin dimers results in the formation of a rigid tube-like structure, which by growing, or polymerizing, or shrinking, depolymerizing, can effect force on a target. Microtubules are involved

in a number of other functions, such as serving as roadways along which transport molecules can carry cargo, and forming the structural backbone of the cell. The spontaneous assembly of tubulin dimers in solution into a microtubule is heavily disfavoured. However, when free floating tubulins encounter pre-formed ring of tubulin molecules, the growth of a microtubule is greatly facilitated. In cells, this template is known as the γ-Tubulin Ring Complex (γ-TuRC) which is a ring-like structure of γ-Tubulin molecules held together by various other proteins known as γ-tubulin ring proteins (GRIPs). γ tubulin shares a similar structure to α and β tubulin and have been shown to act as nucleation templates for microtubules. Furthermore, structural rearrangements of the γ-TuRC have been shown to either facilitate and inhibit nucleation.

1.3.2 γ-Tubulin & the γ-CT

For many years, microtubule nucleation was though to be the sole function of γ tubulin. However, in 2001, a mass spectrometry study of the yeast γ-Tubulin showed that the protein is phosphorylated *in vivo* at 9 sites. Several functional studies following up on the finding that γ-tubulin is regulated show that mutations altering phosphorylation sites, all of which lie in IDRs, have consequences on the organization and stability of microtubules but no effect on their nucleation.

how many sites?

One of the best studied phosphorylation sites in γ-Tubulin is the heavily conserved Tyrosine (Y) 445 which lies in the disordered carboxyl terminal tail of γ-Tubulin (γ-CT). The γ-CT is defined as the final 35 residues in the C-terminal portion of γ-Tubulin, where the folded globular domain ends. Substitution of an Aspartic Acid (D) residue in the place of Y445, making the Y445D mutant, results in

slow growing cells with unstable and misaligned mitotic spindles. The aspartic acid substitution is used as a controllable means for recreating the electrostatic environment of a phosphate group by introducing a negative charge. These findings suggest that γ-Tubulin is playing a role outside of its canonical function of nucleation and is coordinating the dynamics of the mitotic spindle. The coupling of post-translational modifications (PTMs) to the C-terminal tails o tubulins has been described in the γ-tubulin orthologues α and β tubulin. Specific combinations of PTMs on the tubulin tails act as a 'tubulin code'for recruiting motor proteins and microtubule associated proteins to the microtubule lattice. A similar code has not yet been described for γ-Tubulin. While the functional importance of phosphorylation and IDRs in γ-Tubulin is becoming increasingly clear, the physical mechanisms by which local modifications at IDRs can have global impacts on the large molecular machine remain unstudied.

1.4 Experimental question

We hypothesize that the addition of a negative charge at the position Y445 in the γ-CT alters the conformational sampling of the disordered tail in such a was as to be able to regulate the function of the complex.

1.5 Approach

In order to detect the rapid changes in conformational sampling, we use a powerful computational technique known as Molecular Dynamics (MD) simulations. We will be simulating the dynamics of the γ-CT in isolation as well as in the presence of the entire γ-Tubulin protein.

should I bring up NMR here?

Notes from jackie: keep REMD in discussion as possible method. Tony says REMD more

CHAPTER 2
Theory & Methods

> Life can only be understood backwards; but it must be lived forwards.
>
> Søren Kierkegaard

The main technique we will use to study the behaviour of IDPs is the computational technique of Molecular Dynamics (MD) simulations. IDP dynamics are shaped by various types of physical interactions acting on a high number of conformational degrees of freedom all on a very fast timescale. For this reason it is difficult to predict the dynamics of IDPs *ab initio* MD is a brute-force approach which iteratively solves the equations of motion for every interaction in a system of atoms in 3D space. What results is a trajectory in space and a velocity for every atom in the system in time which we can use to visualize the conformational sampling of our IDP of interest and compute thermodynamic quantities. While this can be a computationally demanding task, it is currently the most reliable way of computationally studying the dynamics of molecular systems.

2.1 Molecular Dynamics Simulations

We represent our system as a set of N atoms represented as vectors $R = \{\mathbf{r_1}, \mathbf{r_2}, ..., \mathbf{r_N}\}$ in three dimensional space. We then use classical Newtonian mechanics to obtain the changes in position of the particles as a function of time. For a

peptide in solution, this would consist of the atoms in the peptide, and water atoms and the forces arising from interactions between all the particles in the system.

2.1.1 Computing trajectories of atoms

The central principle of MD is that the potential energy arising from interacting particles is a function of their positions in space. Since force is related to potential energy, it follows that the acceleration of the particles is a function of the potential energy, and so the motion of the particles can be obtained. Given a description of the potentials arising from interactions between the different atoms, which we call a force field, we can iterate through every atom in the system and calculate the force that would arise as a function of the potential energy function. The potential energy given by the force field can be written as $V(\mathbf{r_1}, \mathbf{r_2}, ..., \mathbf{r_N})$ is a function of the positions of each atom. Using the classical definition of force as $\mathbf{F} = m\mathbf{a}$, we can combine the positions of each atom with the force field to compute the force acting on each atom as follows.

$$\mathbf{F_i} = -\frac{\partial V(\mathbf{r_1}, \mathbf{r_2}, ..., \mathbf{r_i}, ...\mathbf{r_N})}{\partial \mathbf{r_i}} \tag{2.1}$$

Given that the force on an atom is the result of interactions with all other atoms in the system, we obtain the force on a particular atom as the sum of the force of the interactions with all other atoms j in the system. So we get $\mathbf{F_i} = \sum_j \mathbf{F_{ij}}$ Given the total force on an atom, we can compute its trajectory in space by numerically integrating Newton's equations of motion. This process is repeated and trajectories are stored and updated for the desired number of steps in the simulation.

$$\frac{\partial^2 \mathbf{r_i}}{\partial t^2} = \frac{m_i}{\mathbf{F_i}} \tag{2.2}$$

2.1.2 Force Field

The functions for potential energy of every type of interaction in the system are defined in what we call a force field. The energy between two interacting atoms can be broken down into two broad types of interactions: bonded and non-bonded interactions.

$$E_{total} = E_{bonded} + E_{nonbonded} \tag{2.3}$$

The bonded energy term can be written as the sum of energies arising from the bond itself (E_{bond})which is a function of the bond length, the potential arising from the angle formed by the bond (E_{angle}), as well as the torsional/dihedral angle ($E_{dihedral}$) arising from the rotation of three bonds about two intersecting planes.

Non-bonded interactions can have two contributing factors; electrostatic force, and van der Waals force. The electrostatic potential ($E_{electrostatic}$) arises from the interaction of the charges of particles, while the van der Waals potential ($E_{vanderWaals}$ arises from the attraction or repulsion between uncharged groups. In most systems, non-bonded interactions by far outnumber bonded interactions and thus carry most of the computational weight in an MD simulation. in m Combining all of these terms, we can write the full description of forces in the system as:

should i include this?

$$E_{total} = E_{bond} + E_{angle} + E_{dihedral} + E_{vanderWaals} + E_{electrostatic} \tag{2.4}$$

$$E(r_N) = \sum_{bonds} K_r(r-r_0)^2 + \sum_{angles} k_\theta(\theta-\theta_0) + \sum_{dihedrals} K_\phi(1+cos(n\phi-\delta)) + \sum_{i,j}\{4\epsilon_{i,j}\frac{\sigma_{ij}}{r_{ij}}\} \tag{2.5}$$

The MD algorithm evaluates $E(r_N)$ at every time step to obtain the force on each atom and therefore the trajectory at each time step. Given such a fine-grain level of modelling, this process is the most time costly step in an MD simulation. However, an important advantage to such a low level description of the system is that more complex phenomena such as the hydrophobic effect and hydrogen bonding which are known to be essential to protein dynamics do not need to be coded explicitly in the models. Instead, they arise naturally from this definition of the system.

Another key component to the force field, is the definition of parameters for the different types of interactions and particles in the system. Parameters can include values for charge, mass, bond length, etc. These are often obtained from experimental measurements. The force field must naturally also contain a set of definitions for the various types of atoms and functional groups it can model. Therefore, the choice of force field can have important consequences on the outcome of the simulations and must be chosen with care.

2.1.3 Preparing the system

The starting point of an MD simulation is a force field and a set of initial coordinates for the system of interest. Before a simulation can be successfully run, there are several pre-processing steps that must be executed.

As mentioned earlier, hydrogen bonding and the hydrophobic effect play very important roles in shaping the dynamics of polypeptides therefore our model must include water molecules. We therefore place the peptide atoms are placed in a

simulated box where water molecules are introduced to fill the remaining space. All subsequent force calculations in MD will consider interactions between solvent-solvent and solvent-peptide atoms. Given that many more water molecules will be present than peptide molecules, computations on solvent atoms are computational costly. Some variations on MD avoid these computations by modelling the solvent implicitly as a mean field instead of explicitly considering every atom in the system.

Once the peptide is solvated any initial steric clashes between atoms must be allowed to relax. Typically this involves executing an energy minimization algorithm which searches for atomic coordinates that minimizes the forces between atoms to move the system towards an energy minimum. No minimization algorithm guarantees convergence to a global minimum in finite time on a realistic system. However, convergence to a local minimum is often sufficient to eliminate significant clashes.

At this point, we could begin an MD simulation and obtain trajectories in the NVE ensemble (constant number of particles, volume, and energy). However, we are often interested in comparing results from MD to experimental measurements such as those from NMR where the system is under constant temperature and pressure. It is therefore necessary to ensure that the forces in the system don't produce large fluctuations in the pressure and temperature of the ensemble. In order to keep the temperature constant and achieve an NVT (constant number of particles, volume, and temperature) we use a thermostat. Since the temperature of a system is a function of the kinetic energy, a thermostat re-scales the velocities of the atoms in the system to achieve a given temperature. Likewise, for maintaining constant pressure, a barostat adjusts the size of the box to counteract fluctuations in pressure

and thus achieving an NPT ensemble (constant number of particles, pressure, and temperature). During the equilibration step, we first let the system equilibrate to the desired temperature by executing a short simulation in NVT. Then under NPT we allow the system to adjust to the desired pressure. Once both equilibration simulations are complete, the system is ready for a full simulation in NPT.

2.2 Trajectory Analysis

The MD simulation generates a set of coordinates for every atom in the system as a function of time, $r(t)$. From these trajectories we can compute several quantities to study conformational changes in the peptide over time.

2.2.1 Root Mean Square Deviation

We measure the square displacement between the coordinates of atom i at time t weighted by the mass of the atom m_i. We iterate this process for every atom in the peptide to obtain a measure of the degree of change between two conformations in time.

$$\mathrm{RMSD}(t_1, t_2) = \left[M^{-1} \sum_{i=1}^{N} m_i ||\mathbf{r_i}(t_1) - \mathbf{r_i}(t_2)||^2 \right]^{\frac{1}{2}} \tag{2.6}$$

2.2.2 Radius of gyration

The radius of gyration is a measure of a structure's compactness. To obtain the radius of gyration, we compute the mean squared distance from every atom r_i to the molecule's centre of mass r_{mean}.

$$R_g(\mathbf{r}) = \sqrt{N^{-1} \sum_{k=1}^{N} (\mathbf{r_k} - \mathbf{r_{mean}})^2} \tag{2.7}$$

2.2.3 Covariance Analysis

When analyzing MD trajectories, we are often interested in observing collective motions. This is because global motions are likely to be involved in some functional mechanism. However, molecular trajectories typically feature complex motions along many axes and time scales which can often make it difficult to detect coordinated motions. For example, local rearrangements, vibrations, rotations, and random diffusion are examples of non-coordinated motions that likely do not contribute to a functional mechanism. The goal in MD trajectory covariance analysis is to obtain the axes of motion where atoms in the peptide of interest show a high degree of correlation which could be indicative of a global coordinated motion.

Covariance analysis, or principal component analysis is a mathematical tool which isolates principal axes, or components of motion by computing the covariance between atoms at every time point in the simulation. We compute the covariance for all N atoms in 3 dimensions, resulting in a covariance matrix of size $3N$.

$$C_{ij} = \left\langle M_{ii}^{\frac{1}{2}}(\mathbf{x_i}(t) - \langle \mathbf{x_i}(t) \rangle) M_{jj}^{\frac{1}{2}}(\mathbf{x_j}(t) - \langle \mathbf{x_j}(t) \rangle) \right\rangle \tag{2.8}$$

The eigenvectors of the covariance matrix, C define the set of orthogonal axes along which maximize variance. Note that $\langle \quad \rangle$ denotes a time average. Due to the constraints imposed by the backbone, only a couple of eigenvectors are expected to contribute most to global movements.

$$R^T C R = diag(\lambda_1, \lambda_2, ...\lambda_{3N}) \qquad \text{where} \quad \lambda_1 \geq \lambda_2 \geq \lambda_{3N} \tag{2.9}$$

Where R is the transformation matrix whose columns contain an eigenvector. Using this matrix to diagonalize C, we get a diagonalized C containing the set of eigenvalues λ_i for every eigenvector in R along its main diagonal. The magnitude of the eigenvalue tells us the amount of variance captured by its corresponding eigenvector and can thus be used to guide our projection toward the major axes of motion.

If we wish to visualize the system's motions along a particular axis and filter out motions along other axes, we can project the coordinates of each atom along an eigenvector. We can the following transformation to obtain the new set of coordinates $\mathbf{p}(t)$.

$$\mathbf{p}(t) = R^T M^{\frac{1}{2}} \mathbf{x}(t) \tag{2.10}$$

The resulting trajectory lets us visualize motions along any component and is a useful tool for detecting coordinated structural changes.

define Rg, Dc, RMSD, contact maps

2.2.4 MD Alternatives

2.3 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[6] 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 [8]. The starting γ-CT polypeptide configurations were obtained from secondary and tertiary structure predictions by RaptorX [7] 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 [9]. Translational diffusion coefficients were calculated for each structure using hydroNMR software [3]. 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.

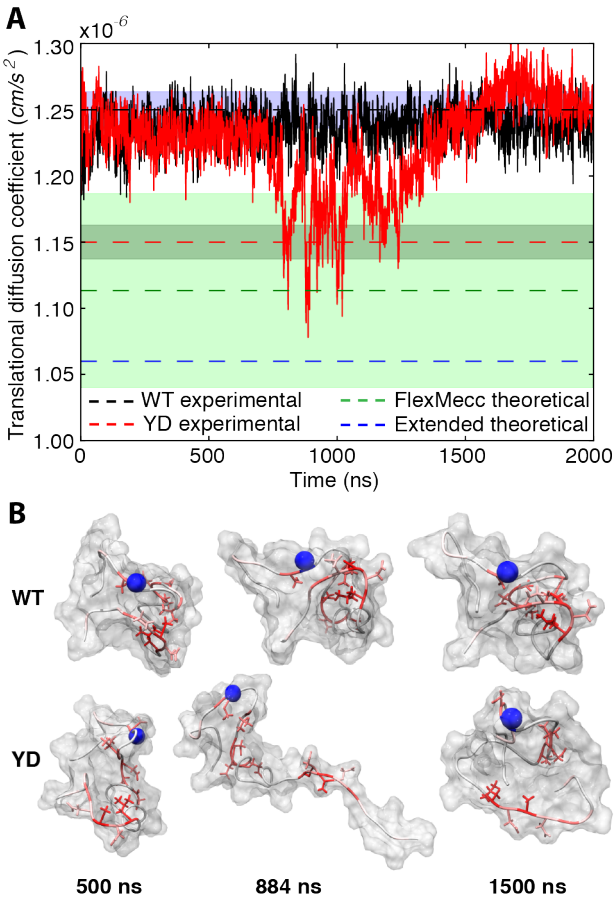


Figure 3–1: Diffusion Coefficient

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.

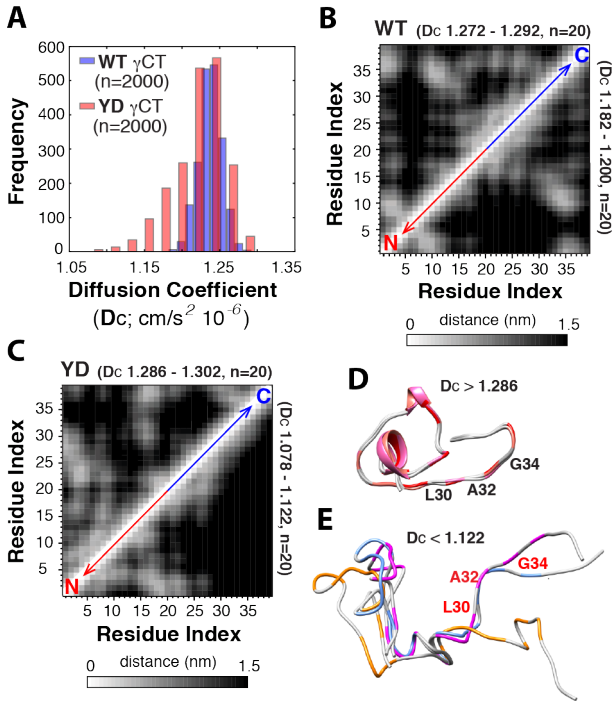
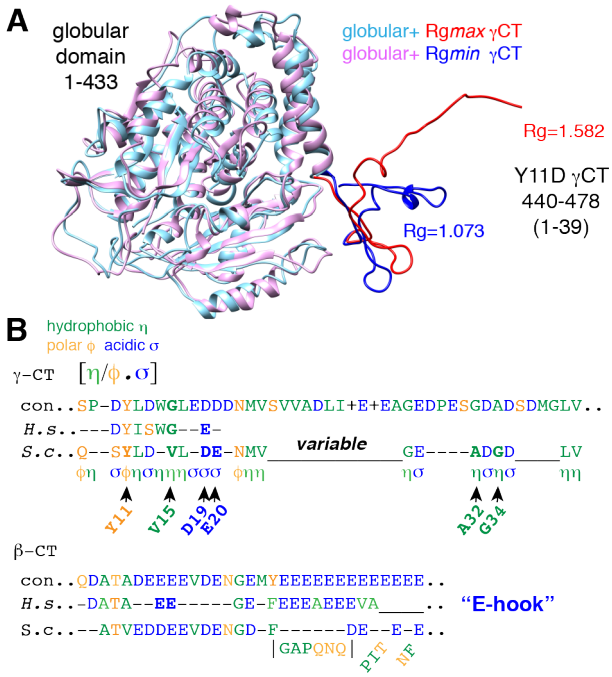


Figure 3–2: Contact Maps

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

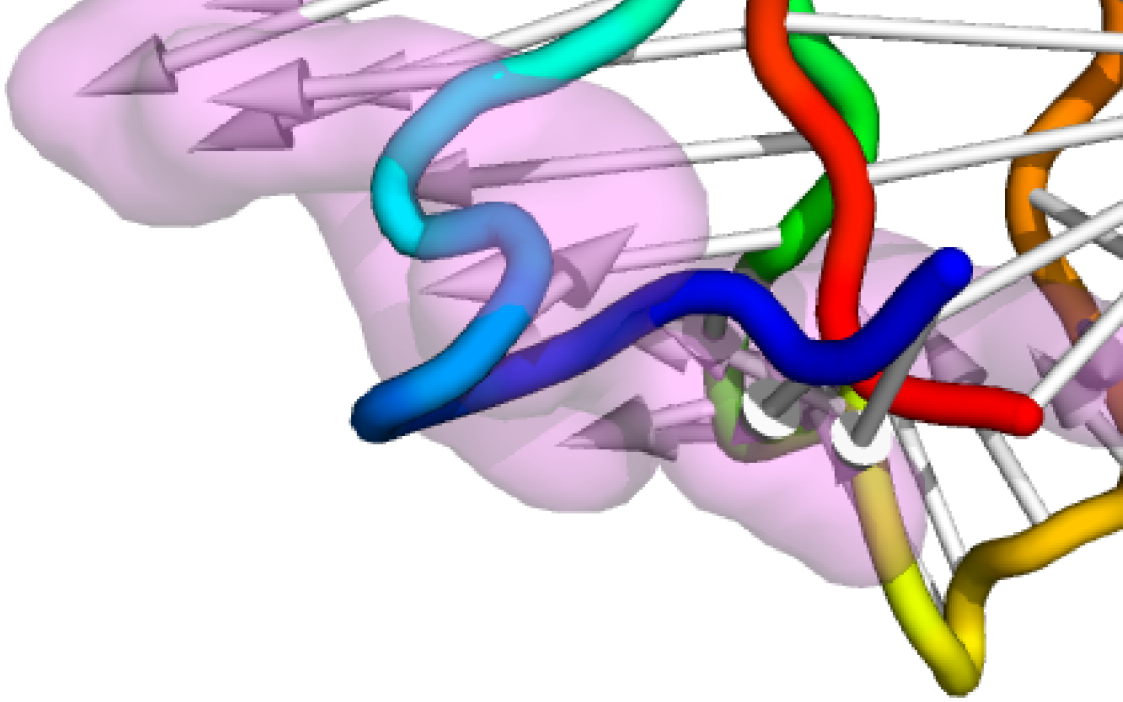
dispersion magnitudes indicated in red. This analysis shows that the extended conformations consist of a compact N-terminus with residues lo-cated in the C-terminal region of the γ-CT, (including dynamical-ly-broadened residues L30, A32 and G34) isolated from the N-terminus and solvent-accessible. Through 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] Roby P Bhattacharyya, Attila Reményi, Matthew C Good, Caleb J Bashor, Arnold M Falick, and Wendell A Lim. The ste5 scaffold allosterically modulates signaling output of the yeast mating pathway. *Science*, 311(5762):822–826, 2006.

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

[3] 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.

[4] Marcos RM Fontes, Trazel Teh, and Bostjan Kobe. Structural basis of recognition of monopartite and bipartite nuclear localization sequences by mammalian importin-α. *Journal of molecular biology*, 297(5):1183–1194, 2000.

[5] Thomas A Graham, Denise M Ferkey, Feng Mao, David Kimelman, and Wenqing Xu. Tcf4 can specifically recognize β-catenin using alternative conformations. *Nature Structural & Molecular Biology*, 8(12):1048–1052, 2001.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] Alexander Sigalov, Dikran Aivazian, and Lawrence Stern. Homooligomerization of the cytoplasmic domain of the t cell receptor ζ chain and of other proteins containing the immunoreceptor tyrosine-based activation motif. *Biochemistry*, 43(7):2049–2061, 2004.

[12] Hongtao Yu, James K Chen, Sibo Feng, David C Dalgarno, Andrew W Brauer, and Stuart L Schrelber. Structural basis for the binding of proline-rich peptides to sh3 domains. *Cell*, 76(5):933–945, 1994.

[13] Tsaffrir Zor, Bernhard M Mayr, H Jane Dyson, Marc R Montminy, and Peter E Wright. Roles of phosphorylation and helix propensity in the binding of the kix domain of creb-binding protein by constitutive (c-myb) and inducible (creb) activators. *Journal of Biological Chemistry*, 277(44):42241–42248, 2002.

KEY TO ABBREVIATIONS