Imperial College London
Department of Mathematics

All-scale structural analysis of biomolecules through dynamical graph partitioning

Antoine Delmotte

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy of Imperial College London, 2014

Copyright

Declaration of Originality

I hereby certify that this thesis and the research to which it refers are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline.

Antoine Delmotte

Abstract

From femtosecond bond vibrations to millisecond domain motions, the dynamics of biomolecules spans a wide range of time and length scales. This hierarchy of overlapping scales links the molecular and biophysical details to key aspects of their functionality. However, the span of scales combined with their intricate coupling rapidly drives atomic simulation methods to their limits, thereby often resulting in the need for coarse-graining techniques which cannot take full account of the biochemical details.

To overcome this tradeoff, a graph-theoretical framework inspired by multiscale community detection methods and stochastic processes is here introduced for the analysis of protein and DNA structures. Using biophysical force fields, we propose a general mapping of the 3D atomic coordinates onto an energy-weighted network that includes the physico-chemical details of interatomic bonds and interactions. Making use of a dynamics-based approach for community detection on networks, optimal partitionings of the structure are identified which are biochemically relevant over different scales. The structural organisation of the biomolecule is shown to be recovered bottom-up over the entire range of chemical, biochemical and biologically meaningful scales, directly from the atomic information of the structure, and without any reparameterisation.

This methodology is applied and discussed in five proteins and an ensemble of DNA quadruplexes. In each case, multiple conformations associated with different states of the biomolecule or stages of the underlying catalytic reaction are analysed. Experimental observations are shown to be correctly captured, including the functional domains, regions of the protein with coherent dynamics such as rigid clusters, and the spontaneous closure of some enzymes in the absence of substrate. A computational mutational analysis tool is also derived which identifies both known and new residues with a significant impact on ligand binding. In large multimeric structures, the methodology highlights patterns of long range communication taking place between subunits. In the highly dynamic and polymorphic DNA quadruplexes,

key structural features for their physical stability and signatures of their unfolding pathway are identified in the static structure.

Acknowledgements

First and foremost, my greatest thanks are for my supervisors, Prof Mauricio Barahona and Prof Sophia Yaliraki. What I learned from them over these past four years is hard to quantify. Much of my scientific mind and my approach to research have been shaped over the course of our discussions over a cup of coffee in one of the local cafes. Their dedication to science and their constant brainstorming for new ideas have been a true inspiration, and the freedom they gave me to try and pursue my own ideas has been invaluable in the accomplishment of my PhD. I am also grateful to them for the opportunity to attend a variety of conferences and summer courses, and most of all, for the numerous collaborations they helped build around my work.

In particular, I had the privilege to work with Prof Michael Thorpe, whose ability to treat the most complex problems with simple tools, along with his intense and fast-paced research style have been extremely enriching to me. Over the course of my first year, I benefited a lot from regular meetings with Prof Renaud Lambiotte, his energy, his contagious enthusiasm, and the wealth of ideas that came from our discussions. I am also thankful to Dr Laura Barter and Dr Rudiger Woscholski for the opportunity to work on one of the most interesting biomolecular structure as well as their frequent and insightful feedback. Without Prof Ramon Villar's open-mindedness and curiosity for novel computational approaches, my work on DNA quadruplexes would never have been possible. Thanks also to Prof Ed Tate for welcoming me into his group for six months and giving me the chance to use my work for the first time to help understand a new protein structure. I should also very much acknowledge Matthew Reynolds whose eagerness for spending his summer holidays in my company to analyse biomolecular structures has been a key contribution to a significant part of this work.

I am much obliged to the British Heart Foundation at Imperial College, as well as Wallonie Bruxelles International for funding my PhD, and giving me the chance to attend numerous workshops and conferences.

Beyond the scientific work, a number of people have made my time at Imperial a truly special experience: in particular my office mates, Mariano Beguerisse-Diaz, Justine Dattani, and Michael Schaub, as well as all the Barahona-Yaliraki group, Benjamin Amor, Elham Ashoori, Elias Bamis, Sarah Byrne, Elisa Dominguez-Hüttinger, Navaneeth Krishnamoorthy, Panayotis Georgiou, Nuno Nene, Neave O'Clery, Zenna Tavares, Borislav Vangelov and Yu William Yun. Thank you all!

Finally, I cannot thank enough my parents Marie-Claire and Patrick, my brother Charles, and, most importantly, Marjolaine, for having always believed in me, and never ceased to encourage me in all my enterprises, whichever they were.

Antoine Delmotte
London, January 2015

Contents

List of Figures

List of Tables

Abbreviations

AdK . . . . . Adenylate kinase
ADP . . . . . Adenosine diphosphate
AMP . . . . . Adenosine monophosphate
Asp . . . . . Apsartate
ATCase . . . Aspartate transcarbamoylase
ATP . . . . . Adenosine triphosphate
CA . . . . . Carbamoyl-aspartate
CABP 4-carboxy-arabinitol 1,5-bisphosphate
CP . . . . . Carbamoyl phosphate
CTP . . . . . Cytidine triphosphate
ELC . . . . . Essential light chain
FRET . . . . Fluorescence resonance energy transfer
GAP45 . . . . 45-kDa glideosome-associated protein
MTIP . . . . Myosin Tail Interacting Protein
MyoA . . . . Myosin A
NMR . . . . . Nuclear magnetic resonance
PALA . . . . N-phosphonacetyl-L-aspartate
PfMTIP . . . *Plasmodium falciparium* MTIP
PGA 3-phosphoglycerate
Pi . . . . . Inorganic phosphate
PkMTIP . . . *Plasmodium knowlesi* MTIP
RLC . . . . . Regulatory light chain
RMSD . . . . Root-mean-square deviation
Rubisco . . . Ribulose 1,5-bisphosphate carboxylase/oxygenase
RuBP Ribulose-1,5-bisphosphate

UTP . . . . . Uridine-triphosphate
VI . . . . . Variation of information
XuBP . . . . Xylulose 1,5-bisphosphate

Chapter 1

Introduction

How does function in biomolecules emerge from their atomic structure? Most biological and physiological processes are the consequence of the action and interplay of a large body of proteins and their ability, arising solely from the properties of their molecular structure, to carry out an immense variety of elementary tasks in the cell. The catalysis of reactions, the transmission of signals inside and outside the cell, the transport of small molecules, the generation of mechanical motion, the provision of structural support, or the control of vital processes such as cell death, DNA replication or the immune response are only some of the vital roles fulfilled by proteins. The understanding of their physical properties is not only crucial to our comprehension of many biological mechanisms, but also to our ability to address issues associated with them. The design of drugs that specifically target a particular protein, RNA molecule, or DNA sequence provides the means to control biological processes, for instance to fight diseases resulting from an infection or anomalies in vital biomolecules. Ultimately, uncovering the design rules that define their behaviour in living organisms opens the door the engineering of biomolecular structures with new or improved functions, acting as biological devices at the atomic level. The link between the chemistry of biomolecules and the ensuing physiological and biological processes is therefore a central question in biology.

Since 1958, when the first atomic resolution structures of proteins, myoglobin and haemoglobin, were published by Kendrew (Kendrew *et al.*, 1958) and Perutz (Muirhead and Perutz, 1963), tremendous progress has been made in experimental techniques and computational modelling. Yet, due to the complex structural properties of biomolecules, predicting their dynamical behaviour and rationalising the mechanisms that give rise to their function remain real challenges more than

five decades later. Specifically, the structural and dynamical aspects that lead to their macroscopic properties span an extensive range of scales and evolve from the atomistic level up.

From a structural viewpoint, biomolecules are characterised by a multi-level organisation. Each protein, DNA or RNA molecule is formed of one or more chains of amino or nucleic acid residues whose sequence, called the primary structure, uniquely encodes the biomolecule. Locally, the residue chains often arrange into particular structural motifs, referred to as the secondary structure, such as helices or pleated sheets in proteins. Globally, biomolecules fold into a specific three dimensional conformation, the tertiary structure, which in proteins defines domains with a functional role. At the highest level of organisation, multiple chains can associate to produce a so-called multimeric protein that is described by its quaternary structure.

But biomolecules are not static entities. Their biological function often depends on their ability to transition between multiple structural configurations. The binding and release of substrates, the optimal positioning of bound molecules to catalyse a reaction, or the creation of mechanical motion, to cite only a few, all depend on the ability of the protein to change the arrangement of its atoms under particular circumstances. The constraints imposed by the structure, which for instance define rigid and flexible regions, in turn determine the subspace of possible displacements. This ensemble of conformations, and the probability with which they are sampled by the biomolecule is thus itself encoded by the structure, and this synergy between structure and dynamics is at the very heart of its function (Henzler-Wildman *et al.*, 2007a; Henzler-Wildman and Kern, 2007).

The difficulties in predicting the behaviour of biomolecules notably result from the multiple scales, in both time and space, over which their structure and dynamics unfold (Henzler-Wildman *et al.*, 2007a; Henzler-Wildman and Kern, 2007; Frauenfelder *et al.*, 2001; Yaliraki and Barahona, 2007). Atoms, functional chemical groups, amino acids, the ensuing secondary structures, the large conformational domains: each of these levels of structural organisation is linked to dynamics occurring at different time and spatial scales, from the femtosecond vibration of covalent bonds to the micro to millisecond motion of functional domains (Figure 1.1).

The large scale domain motions are often tightly linked to function, and have consequently long been the subject of intense research efforts in biochemistry. Yet the seemingly random atomic fluctuations are not devoid of functional significance. As suggested by recent experimental data combined with the success of computational methods such as normal mode analysis, they instead actively drive the

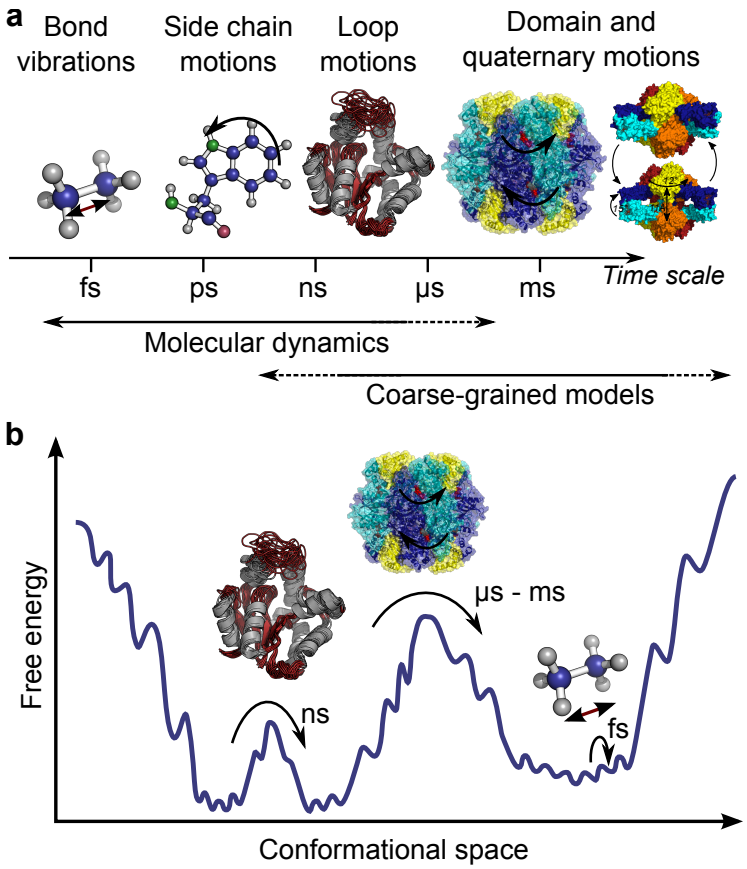


Figure 1.1: Proteins are complex structures with a multiscale organisation in time and space. Fully atomic simulations cannot access the large scales while coarse-grained models cannot fully describe the atomic details of the biochemistry. **a.** Time scales of motions in biomolecules. **b.** Cross-section through the free energy landscape of protein conformations, showing the energy barriers associated with motions over different time scales. The global dynamical behaviour of biomolecules results from the intricate coupling of the dynamics taking place over a deep hierarchy of scales (Henzler-Wildman and Kern, 2007).

collective motions. In the protein adenylate kinase, a hierarchy of dynamics taking place over different time and spatial scales was notably identified whereby motions, even at the smallest scales, occur preferentially along the trajectory towards the catalytically competent conformation (Henzler-Wildman *et al.*, 2007b). Local fast time scale dynamics are thus directly facilitating their global slow time scale counterpart. In between the atomic and macroscopic scales, a deep hierarchy of levels of organisation thus exists which are not behaving independently but rather influence each other and, through their coupled dynamics, all contribute towards the global behaviour of the entire biomolecule.

The range of time and spatial scales spanned by the dynamics of biomolecules presents a real challenge for computational methods such as molecular dynamics simulations. The ability to accurately reproduce the dynamics of molecules computationally from the numerical integration of the Newton's equations of motion combined with meticulously designed interatomic potential energy functions has been a remarkable success of computational chemistry (Adcock and McCammon, 2006). However, simulating the atomic motions of a whole protein over the biologically relevant regimes would generally require spanning ten to fifteen orders of magnitude in time, which would rapidly drive such fully atomic molecular dynamics simulation methods to the limits of available computational resources. Thus, although traditional tools can deal successfully with the very short time scales, they often cannot be applied to long times (or large systems) due to their exorbitant computational cost.

Because many of the key biological functions take place at the micro- to millisecond time scales, simplified structure models using coarse-graining techniques have been proposed as a means to reaching the biologically relevant regimes (Bahar and Rader, 2005; Tozzini, 2005; Ayton *et al.*, 2007). However, in addition to ignoring the details of physico-chemical atomic interactions, coarse-grained models also effectively decouple the smaller from the larger levels of structural organisation. Consequently, they are usually unable to link atomic scale events such as substrate binding with the large-scale conformational changes induced and cannot provide a picture that emerges seamlessly from the smallest scales.

The different levels of organisation in proteins do not indeed behave independently: the dynamics at long time and length scales, which is in many cases crucial for biological function, is the result of the integrative interaction of the finer organisational levels. Analysing proteins from this multiscale perspective can reveal the intricate linkage between the structural levels of organisation and give insight into the behaviour of the protein starting from the bottom-up. In addition, this picture can also aid in understanding the effects that small-scale changes such as mutations have on the large-scale behaviour.

1.1 Organisation of the thesis

To address this problem, a new computational approach is here introduced which provides the means to explore the structural organisation of a protein, DNA or RNA molecule in relation to its dynamical behaviour throughout the entire spectrum of

scales. Importantly, this is done using its static X-ray or NMR structure only, defined at the atomic level, and without the use of any coarse-graining, *a priori* information or assumption.

While simulations are rapidly limited by computational resources, it is however well established that the static structure of the protein encodes to a large extent the space of accessible motions which, in turn, provides the mechanisms allowing the protein to perform its specific function. Normal mode analysis (Brooks and Karplus, 1983; Go *et al.*, 1983), elastic network models (Bahar *et al.*, 2010), graph theoretical approaches (Csermely *et al.*, 2013; Di Paola *et al.*, 2013; Böde *et al.*, 2007), and rigidity theory (Costa, 2008; Jacobs *et al.*, 2001) are some examples of the many computational methods which have successfully made use of these conclusions to infer properties of the protein dynamics and function, and we review and discuss each of them in Chapter 2.

Our methodology is described in Chapter 3. It is twofold: Firstly, we define a general mapping to convert a fully atomic biomolecular structure into a network of atoms that includes all the physico-chemical details of the interactions; Secondly, we use a dynamics-based approach for multiscale graph partitioning (Delvenne *et al.*, 2010) which uncovers graph communities that are relevant over different time scales using a stochastic process diffusing on the graph.

This leads to a multi-level hierarchical organisation of the biomolecular structure that identifies the biochemically meaningful substructures at all scales: from chemical groups through individual residues, to the appearance of secondary structures and intermediate structural elements, such as clusters of several helices, to the eventual emergence of large conformational units. Hence the picture at larger scales emerges directly from the detailed physico-chemical information at the smallest atomic scales. These results are exemplified in Chapter 4 on adenylate kinase, a classical and well studied enzyme, where we also define two biochemically motivated surrogate random graph models that allow us to evaluate the significance of our results and identify the role of particular types of bonds and interactions.

In Chapters 5 to 7, this general framework is then used to understand the multiscale dynamical features and infer possible mechanisms of functional motions of an ensemble of increasingly complex protein and DNA structures which are currently actively researched. In Chapter 5, we identify in the myosin tail interacting protein, a particular myosin-myosin light chain interaction from the malaria parasite, regions that share a common dynamical behaviour, such as a rigid cluster and regions with a common functional role. We also introduce a computational mutational analysis

tool that evaluates the impact of individual residues on the global structural organisation of the protein. In Chapter 6, we demonstrate the ability of our methodology to provide insight into highly complex multimeric structures such as Rubisco and ATCase. We show that the complexity of their functional mechanisms is encoded in their structure in terms of a deep hierarchy of functionally important levels of organisation which contrasts with simpler globular structures such as hemoglobin. Through the comparison of different reaction stages and a methodology to analyse the ensemble of suboptimal solutions, we reveal particular patterns of communication taking place at the level of the quaternary structure, throughout the multimer. In Chapter 7, we extend our methodology to the analysis of DNA structures. Our study of 19 DNA quadruplexes highlights particular structural features related to the stability, polymorphism, and unfolding pathway of the different structures.

Chapter 2

Computational analysis of biomolecular structures

A wide range of techniques now exist which provide the means to observe experimentally different aspects of the structure and dynamics of biomolecules. Their three dimensional structure can notably be resolved at the atomic level using nuclear magnetic resonance (NMR) spectroscopy and X-ray crystallography, or at a lower resolution using cryo-electron microscopy. Today, more than 100 000 structures can readily be accessed in the Protein Data Bank, the largest such database (Bernstein *et al.*, 1978). In addition, NMR spectroscopy also allows to probe the dynamics over a broad range of time scales, from pico- to millisecond motions, and to infer the probability of the different conformations together with the rate of transitions between them. The motion of a single molecule can even be followed in real time using fluorescence resonance energy transfer (FRET) and normal modes of motions can be identified using infrared (IR) spectroscopy.

Yet experiments remain limited in the level of detail they can provide on the dynamics of biomolecules. For instance, short-lived transition conformations are usually inacessible and the exact pathway between different well defined states remains out of reach to experimentalists. As computational methods do not suffer from this limitation, they can reveal key functional properties that remain unobservable experimentally. Simulations and computational analyses of experimental structures have for instance been used to discriminate between possible functional mechanisms, identify residues critical for function, suggest preferential communication pathway accross the structure, characterise allosteric mechanisms, decompose the structure into rigid clusters, or again uncover new binding sites.

Computational methods can thus both complement the information gathered experimentally and anticipate the consequences of particular biochemical events to help guide future experiments. Used in synergy with experiments, they open the door to an enhanced understanding of protein structure and dynamics. Over the past decades, a wealth of methods have been proposed and in the following sections, we describe popular approaches related to this work.

2.1 Normal mode analysis

2.1.1 General approach

In mechanical systems, motions describing small deviations around a stable equilibrium position can often be very well approximated by a superposition of independent vibrational motions. This well-known process of decomposition into decoupled harmonic oscillations has a long history and is of widespread applicability in a broad range of physical systems (Goldstein, 1953). In chemistry, it has long been a very popular method, and was originally used for small molecules, in connection with the vibrational spectrum observed experimentally in Raman and infrared spectroscopy (Wilson *et al.*, 1955). Since the 1980's and its first application to proteins (Brooks and Karplus, 1983), it has become a standard method to analyse the dynamics of large biomolecules. Its interest lies in its ability to reveal collective fluctuations spanning the whole system, and the recent realisation that low frequency modes are often related to the biological function (Ma, 2004, 2005). As a result, normal mode analysis (NMA) is now primarily used to study large scale motions and slow time scale dynamics. In this section, we first describe the mathematical basis of the methodology and then discuss its uses and limitations.

Let us consider a molecule containing N atoms characterised by 3N cartesian coordinates. For an atom i, the spatial coordinates are represented by a vector $\boldsymbol{r}_i = (x_i, y_i, z_i)$, and its displacement relative to a reference position $\boldsymbol{r}_{i0}$ by the vector $\boldsymbol{q}_i = \boldsymbol{r}_i - \boldsymbol{r}_{i0}$. Using a series expansion, the potential energy V of the molecule around the reference point $\boldsymbol{r}_{i0}$ can be written

$$V(\boldsymbol{q}) = V_0 + \sum_{\alpha} \sum_{i=1}^{\mathrm{N}} \left.\frac{\partial V}{\partial q_i^{\alpha}}\right|_{\boldsymbol{q}=\boldsymbol{0}} q_i^{\alpha} + \frac{1}{2} \sum_{\alpha,\beta} \sum_{i,j=1}^{\mathrm{N}} \left.\frac{\partial^2 V}{\partial q_i^{\alpha} \partial q_j^{\beta}}\right|_{\boldsymbol{q}=\boldsymbol{0}} q_i^{\alpha} q_j^{\beta} + \Theta(\boldsymbol{q}^3), \qquad (2.1)$$

where the superscripts α and β extend over the x, y and z components of the vectors of coordinates and $\boldsymbol{q}$ is the 3N×1 vector $(q_1^x, q_1^y, q_1^z, q_2^x, q_2^y, ...)^T$. By choosing the reference point at the energy minimum, the second term of Equation 2.1 becomes equal to zero and, since the reference potential can be chosen arbitrarily, the first term can be set at $V_0 = 0$. If the atomic fluctuations are small in amplitude, the higher order terms can be neglected and V can be approximated by a harmonic potential

$$V(\boldsymbol{q}) = \frac{1}{2} \sum_{\alpha,\beta} \sum_{i,j}^{\mathrm{N}} \left. \frac{\partial^2 V}{\partial q_i^\alpha \partial q_j^\beta} \right|_{\boldsymbol{q}=\boldsymbol{0}} q_i^\alpha q_j^\beta . \tag{2.2}$$

Using the Euler-Lagrange formulation of the equations of motion $\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial T}{\partial \dot{q_i}^\alpha} + \frac{\partial V}{\partial {q_i}^\alpha} = 0$, with the kinetic energy $T = \frac{1}{2}\sum_\alpha \sum_{i=1}^{\mathrm{N}} \mathrm{m}_i \left(\dot{q_i^\alpha}\right)^2$ and m_i the mass of atom i, the motion of the molecule can be written in matrix form as

$$\boldsymbol{H}\boldsymbol{q} = \boldsymbol{M}\ddot{\boldsymbol{q}} \tag{2.3}$$

where $(\boldsymbol{H})_{ij}^{\alpha\beta} = -\left. \frac{\partial^2 V}{\partial q_i^\alpha \partial q_j^\beta} \right|_{\boldsymbol{q}=\boldsymbol{0}}$ is the 3N×3N Hessian matrix of the potential V (with a minus sign), and $(\boldsymbol{M})_{ij} = m_i \delta_{ij}$ is the diagonal matrix of atomic masses. Using the mass-weighted Hessian $\boldsymbol{H'} = \boldsymbol{M}^{-\frac{1}{2}}\boldsymbol{H}\boldsymbol{M}^{-\frac{1}{2}}$ and the mass-weighted coordinates $\boldsymbol{q'} = \boldsymbol{M}^{\frac{1}{2}}\boldsymbol{q}$, Equation 2.3 becomes

$$\boldsymbol{H'}\boldsymbol{q'} = \ddot{\boldsymbol{q}}'. \tag{2.4}$$

Since $\boldsymbol{H'}$ is a symmetric matrix, it can be diagonalised by a unitary matrix $\boldsymbol{A}$, with $\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{I}$, such that

$$\boldsymbol{H'} = \boldsymbol{A}^T \boldsymbol{\Lambda} \boldsymbol{A} \tag{2.5}$$

where $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues of $\boldsymbol{H'}$, and $\boldsymbol{A}$ the matrix of the normalised eigenvectors of $\boldsymbol{H'}$. Using a second change of coordinates $\boldsymbol{Q} = \boldsymbol{A}\boldsymbol{q'}$, Equation 2.4 reduces to 3N independent differential equations

$$\ddot{Q}_k(t) = \lambda_k Q_k(t), \quad k = 1, ..., 3N \tag{2.6}$$

whose solution

$$Q_k(t) = c_k \cos(\omega_k t + \phi_k), \quad k = 1, ..., 3N \tag{2.7}$$

defines a vibrational motion with frequency $\omega_k = \sqrt{\lambda_k}$ given by the square root of the k^{th} eigenvalue of the Hessian matrix H. The amplitude c_k and the phase-shift ϕ_k are determined by the initial conditions and, when the system is at thermodynamic equilibrium[1] $c_k = \sqrt{k_b T / w_k^2}$, with k_b being the Boltzmann constant and T, the temperature in Kelvin (Hayward, 2001). Unlike Equation 2.3, the motion associated with each coordinate $Q_k(t)$ in Equation 2.7 is now independent from all the other coordinates. The changes of coordinates have thus decomposed the motion of the molecule into 3N oscillations which are orthogonal to each other, i.e. none of the motions described by any of the modes $Q_k(t)$ can be reproduced by any combination of the others. These vibrations defined by the coordinates $Q_k(t)$ are the *normal modes* of the molecule. As a side note, only $3N-6$ normal modes actually correspond to pure vibrations in molecules. Six modes will indeed always have a zero eigenvalue, and correspond to the six translational and rotational degrees of freedom of the rigid body (Wilson *et al.*, 1955).

By reverting back to the atomic coordinates, the motion $\boldsymbol{q}(t)$ of each atom of the molecule can now be expressed as a combination of normal modes $Q_k(t)$

$$\boldsymbol{q}(t) = \sum_{k=1}^{3N} c_k \boldsymbol{a'}_k \cos(\omega_k t + \phi_k), \tag{2.8}$$

where $\boldsymbol{a'}_k$ is the k^{th} mass-weighted eigenvector of $\boldsymbol{H'}$ (k^{th} column of $\boldsymbol{M}^{-1/2}\boldsymbol{A}$). The contribution of each mode to the motion of each atom is thus determined by the eigenvector of the Hessian matrix, and the mass of each atom.

In summary, under the assumption that the potential can be approximated by a harmonic well at the energy minimum, normal mode analysis (NMA) decomposes the motion of a molecule into a sum of independent modes of vibration at different frequencies. Each mode is associated with one eigenvalue of the Hessian matrix, which determines the frequency of the vibrations ω_k, and one eigenvector $\boldsymbol{a}_k$, which dictates its contribution to the motion of each atom.

Normal mode analysis is useful for computing a variety of properties in biomolecules such as the mean square displacement of atoms, motional correlations, several thermodynamic quantities as well as the vibrational spectrum which can be related to experimental data such as infrared or Raman spectroscopy (Brooks *et al.*, 1988; Wilson *et al.*, 1955). NMA can also be used to improve sampling in MD simulations (Bahar and Rader, 2005; Hayward and de Groot, 2008), or refine experimental struc-

[1] This can be shown using the equipartition theorem, i.e. assuming that each mode contributes equally to the total energy.

tures (Ma, 2004). Indeed, atomic fluctuations can be directly computed from the normal modes. From equation (2.8) with $c_k = \sqrt{\frac{k_b T}{w_k^2}}$, and assuming thermodynamic equilibrium, the mean square displacements are given by

$$\left\langle q_l^2 \right\rangle = \frac{k_b T}{m_l} \sum_{k=1}^{3N} \left(\frac{a_{lk}}{\omega_k} \right)^2 , \tag{2.9}$$

where a_{lk} indicates the l^{th} component of the k^{th} column of the matrix $\boldsymbol{A}$. Through the presence of the freqency ω_k in the denominator, this expression shows that the lowest frequency modes are the dominant contributors to the atomic fluctuations and therefore drive the largest molecular motions. In addition, these slow modes are often found to be highly delocalised and can thus reveal collective motions engaging large portions of the structure (Brooks and Karplus, 1983).

Over the past years, there has been an increasing body of evidence that slow modes also often bear functional significance. Functionally related conformational transitions were found in multiple proteins and complexes to follow one or several of the slow modes (see reviews by Tama and Sanejouand (2001); Ma (2004, 2005), and references therein). This observation has led to a new surge of interest in normal mode analysis, with a multitude of new algorithms and NMA-related methods being proposed, and NMA has now become a common technique to predict large structural deformations related to function.

Despite its agreement with experimental data, NMA has often been criticised for being used beyond the theoretical limits of its validity. Because it relies on the assumption of small deviations from the energy minimum, it is indeed ill-suited to study large conformational changes. A large number of constraints, such as steric hindrance, will for instance be violated by all but the smallest excursions from the equilibrium position. At physiological temperatures, large biomolecules evolve on a very rugged energy landscape with multiple local minima and energy barriers of various heights (Frauenfelder *et al.*, 1991; Henzler-Wildman and Kern, 2007), in contrast to the smooth harmonic potential assumed by NMA. A common alternative is to infer the Hessian from the fluctuations measured in short molecular dynamics simulations. This procedure, known as quasi-harmonic dynamics (Levy *et al.*, 1984), thereby allows to take some account of the anharmonicity. Essential dynamics (Hayward and de Groot, 2008; Ichiye and Karplus, 1991; Amadei *et al.*, 1993), a very similar method, uses principal component analysis (another eigenvalue problem) on the correlation matrix of atomic fluctuations obtained from molecular dynamics trajectories. Essential dynamics thus filters out the major collective modes

from the local fluctuations to identify the main, functionally important, directions of motions.

NMA can also be very demanding in computational resources, firstly, because of the computationally expensive energy minimisation required to bring the molecule in the equilibrium conformation and, secondly, because of the diagonalisation of a very large 3N×3N Hessian matrix. As a result, even though a number of methods have been proposed to reduce the computational cost such as computing the Hessian in the dihedral angle space, or coarse-graining it into rigid blocks (Hayward, 2001), the calculation of NMA remains limited to relatively small proteins. Finally, the computation of the Hessian can be sensitive to the accuracy of the energy minimisation and the choice of empirical force fields used to evaluate the atomic potential function.

2.1.2 Elastic Network Models

The realisation of the functional importance of slow modes combined with the computational cost associated with normal mode analysis led to the development of a multitude of new methods for simplified or coarse-grained NMA. Recent advances were especially sparked by the seminal work of Tirion (Tirion, 1996) who demonstrated that slow modes are extremely robust with respect to the atomic details of the interactions, rendering the accurate definition of the atomic potentials and their second derivative superfluous. Elastic networks models, inspired by Tirion's work, are probably one of the most popular classes of simplified NMA methods for proteins.

The Hessian matrix used in NMA, which can equivalently be seen as the matrix of spring constants associated with the harmonic potentials, is the only input required to compute the normal modes. In order to avoid the expensive computation of this whole matrix from detailed empirical potentials, Tirion proposed the use of a unique spring constant k for all pairs of atoms within a chosen distance cut-off, thereby defining a potential function with only a single parameter to be determined,

$$V_{\mathrm{ENM}}(\boldsymbol{q}) = \frac{k}{2}\sum_{i<j}^{\mathrm{N}}\left(\left\|\boldsymbol{r_{ij}}\right\| - \left\|\boldsymbol{r_{ij0}}\right\|\right)^2 \tag{2.10}$$

where $\boldsymbol{r}_{ij} = \boldsymbol{r}_i - \boldsymbol{r}_j$ and the sum is done over all pairs of atoms i and j within a certain cut-off r_c (such that $\left\|r_{ij0}\right\| < r_c$), chosen by Tirion between 1.1 and 2.5Å. In addition, she assumed this expression to be valid for any conformation, thus avoiding

the lengthy energy minimisations usually required in NMA. Using a second order expansion of equation (2.10), the potential used by Tirion can be written

$$V_{\mathrm{ENM}}(\boldsymbol{q}) = \frac{k}{2}\sum_{i<j}^{\mathrm{N}}\left[\left(\boldsymbol{q_i} - \boldsymbol{q_j}\right).\boldsymbol{u_{ij0}}\right]^2 \tag{2.11}$$

where $\boldsymbol{u_{ij0}} = \frac{\boldsymbol{r_{ij0}}}{\left\|\boldsymbol{r_{ij0}}\right\|}$, which gives a 3N×3N Hessian matrix in the form of a N×N matrix where the element $H_{i,j}$, $i, j = 1, ..., N$ is a 3×3 matrix defined by

$$(\boldsymbol{H}_{\mathrm{ENM}})_{i,j} = \begin{cases} -k.\begin{pmatrix} \left(u_{ij}^x\right)^2 & u_{ij}^x.u_{ij}^y & u_{ij}^x.u_{ij}^z \\ u_{ij}^y.u_{ij}^x & \left(u_{ij}^y\right)^2 & u_{ij}^y.u_{ij}^z \\ u_{ij}^z.u_{ij}^x & u_{ij}^z.u_{ij}^y & \left(u_{ij}^z\right)^2 \end{pmatrix} & \text{if } i \neq j \text{ and } \|\boldsymbol{r}_{ij0}\| < r_c \\ -\sum_{j\neq i}(\boldsymbol{H}_{\mathrm{ENM}})_{i,j} & \text{if } i = j \end{cases} \tag{2.12}$$

Only a single parameter, the spring constant k, thus needs to be fitted to experimental data. This is usually done by adjusting k such that atomic fluctuations estimated from Equation 2.9, where k is simply a scaling factor, best reproduce the atomic mobility experimentally measured in crystallographic B-factors. In spite of the very simplistic form of this potential, Tirion showed that the results of the normal mode analysis were very close to those computed using detailed empirical force fields and preceded by an energy minimisation (Tirion, 1996). Her conclusions have since been confirmed by the many simplified models inspired by her work, most of which had the capacity to reproduce the slow modes obtained from detailed empirical force fields in very little computational time (Bahar *et al.*, 2010). In particular, elastic network models coarse-grained at the residue level (Atilgan *et al.*, 2001; Chennubhotla *et al.*, 2005; Doruker *et al.*, 2000; Micheletti *et al.*, 2004) (sometimes referred to as the anisotropic network model) with distance cut-off usually between 7 and 10Å (Hayward and de Groot, 2008), or non-uniform spring constants (Hamacher and McCammon, 2006; Lyman *et al.*, 2008; Bongini *et al.*, 2010; Hinsen, 1998) have been proposed which demonstrated good agreement with molecular dynamics simulations, NMR or X-ray experimental data.

2.2 Graph theoretical methods

Euler's solution to the problem of the seven bridges of Königsberg (Euler, 1736), published in 1736, is often referred to as the first article on graph theory. Many mathematical properties of graphs are now well understood, making graph theory an invaluable way to model a variety of systems spanning a broad range of disciplines such as the Internet, electric power grids, transportation, metabolic networks, protein-protein interactions, neural networks, social interactions, or scientific collaborations, to mention only a few. More recently, the increase in the availability of relational data and computational resources has triggered a new surge of interest in graph theory and the study of complex networks (Boccaletti *et al.*, 2006). Revealing particular patterns of interactions, critical elements for function, or obtaining a meaningful coarse-grained description of a complex system are some examples of problems where graph theory can provide valuable insight.

Graphs, also often refered to as networks, are abstract structures defined simply by a collection of nodes and a collection of edges which specifies the connections between the nodes. A common form of representation of a graph with N nodes is the N×N adjacency matrix $\boldsymbol{A}$ such that

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge between node i and node j} \\ 0 & \text{otherwise} \end{cases}$$

When the strength of the connection between the nodes matters, a graph can also be weighted, in which case the elements of the adjacency matrix $A_{ij} \in \mathbb{R}^{\mathrm{N}\times\mathrm{N}}$ can take any real positive value. Similary, the relation between the nodes can be bi- or unidirectional (i.e. i is connected to j, but j is not connected to i). When the graph contains no unidirectional links, it is said to be undirected and its adjacency matrix is symmetric.

The degree of a node is defined as the total number of edges it is associated with. In a weighted network, the weighted degree, or strength, k_i of a node is the sum of the weights of its connections to other nodes[2] $k_i = \sum_j A_{ij}$. The Laplacian matrix is another important matrix in graph theory and is defined as

$$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A} \tag{2.13}$$

[2]In the rest of this work, only weighted networks will be considered and we will often use the term *degree* in place of *weighted degree*.

where $\boldsymbol{D}$ is the diagonal matrix of the weighted degrees, $D_{ij} = k_i\delta_{ij}$.

2.2.1 Graph theoretical analysis of protein structures

Following from the growing popularity of networks and complexity science, a variety of methods have been proposed to analyse protein structures from a graph theoretical viewpoint. Biomolecules can indeed intuitively be rationalised as networks of atoms or residues interconnected by an ensemble of physico-chemical bonds and interactions and, unsurprisingly, a considerable range of tools for network analysis have been applied to biomolecules.

For instance, clusters of tightly connected residues, or "communities", in the protein graph can be expected to have a similar behaviour. Remarkably, communities in the protein graph have consistently been found to correspond to protein functional domains using different techniques such as spectral graph partitioning (Kundu *et al.*, 2004; Kannan and Vishveshwara, 1999) or the minimum cut (Xu *et al.*, 2000). The network representation has also proved to be valuable for the analysis of long-range communication mechanisms within the proteins, such as allostery[3]. Path of sequentially connected residues can intuitively been thought of as possible communication channels that could propagate perturbations via residue-residue interactions, and random walks on graphs provide a simple and intuitive model for signal transmission between residues. Consequently, residues with a high centrality[4] (Vendruscolo *et al.*, 2002; Amitai *et al.*, 2004; del Sol and O'Meara, 2005; del Sol *et al.*, 2006), located at the interface between network communities (Chennubhotla and Bahar, 2006; del Sol *et al.*, 2007; Sethi *et al.*, 2009), or impacting random walks on the graph (Lu and Liang, 2009; Park and Kim, 2011) have been repeatedly linked to allostery, protein folding, or function, or used to directly identify active site or binding hotspot residues. Bridging these different concepts, Chennubhotla and Bahar (2006) used the stationary distribution of a Markov process diffusing on the network of residues (which can be equivalently defined as a random walk) to derive a hierarchical soft partitioning of the protein graph[5]. Based on the community ownership patterns of the nodes, they classified residues by their role in the transmission

[3]Mechanism by which the activity of some proteins is regulated through the binding of molecules, called the effector, to an allosteric binding site usually distinct, and sometimes far away, from the protein active site.

[4]Traversed by a large number of (shortest) paths between pairs of nodes in the graph

[5]A division of the network into communities where each node is associated with a probability of belonging to each community.

of allosteric signals: "hubs" are residues with a strong ownership to one particular community which could "broadcast" the signal, and "messengers" are residues shared by several communities which could efficiently transfer information from one group of residues to another. In a related work, Lu and Liang (2009) used the time-dependent response of localised Markov perturbations as they evolve on the network to characterise residues by their efficiency in propagating inter-residue signals.

Other graph theoretical concepts, for instance linked to the number of closed walks or spectral properties, have been used to characterise the degree of folding (Vendruscolo *et al.*, 2002; Estrada, 2000), or disorder (Csermely *et al.*, 2012) of protein structures. Protein structure networks have also been analysed through spectral graph theory (Vishveshwara *et al.*, 2002), notably to identify key residues for protein-protein association at subunit interfaces (Brinda *et al.*, 2002) and clusters of residues linked to folding intermediates (Kannan and Vishveshwara, 1999). Hubs (nodes with high degree) were also suggested as possible mutation sites to alter protein thermal stability (Brinda and Vishveshwara, 2005).

Protein networks have also been characterised with respect to a variety of network properties such as average shortest path, clustering coefficient, assortativity and degree distribution (del Sol and O'Meara, 2005; Böde *et al.*, 2007; Greene and Higman, 2003; Atilgan *et al.*, 2004). Many of these properties are however tightly linked to the way the graph has been constructed from the protein structure and, although the use of a distance cut-off for the assignment of edges is often the preferred method, different strategies can be considered depending on the biophysical properties to be studied as we will discuss in the next chapter.

In the next sections, we focus on two popular methods, namely the Gaussian network model and rigidity analysis, which are more closely related to our methodology and establish a link between graph theoretic concepts and protein dynamics.

2.2.2 The Gaussian Network Model

The Gaussian Network Model (GNM) was originally proposed by Bahar *et al.* (1997) as a new type of elastic network model. It offers a further simplification of the 3N×3N fully atomic Hessian matrix into the N'×N' (where N' is the number of residues) adjacency matrix of the residue contact graph using two fundamental assumptions: residue fluctuations are assumed to be, firstly, Gaussian and, secondly,

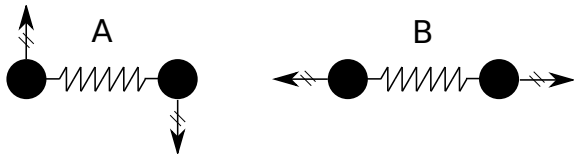


Figure 2.1: According to the GNM potential, motions A and B yield the same change in the potential energy (Thorpe, 2007).

isotropic. In addition, the model is coarse-grained at the level of the residues, using the C_α atoms as the nodes of the network[6].

The GNM is based on the same principles as the classical elastic network models, using this time a potential defined by

$$V_{\mathrm{GNM}} = \frac{k}{2} \sum_{i<j}^{N} \left\| \boldsymbol{r_{ij}} - \boldsymbol{r_{ij0}} \right\|^2 = \frac{k}{2} \sum_{i,j}^{N} \left\| \boldsymbol{q_i} - \boldsymbol{q_j} \right\|^2 \tag{2.14}$$

where the sum extends here over all pairs of C_α atoms i and j within a fixed distance cut-off r_c ($\left\| \boldsymbol{r_{ij0}} \right\| < r_c$). Unlike the classical elastic network models, the GNM potential is not a function of the distance between the two atoms connected by a spring, but of the norm of the vector describing the transition from $\boldsymbol{r}_{ij0}$ to $\boldsymbol{r}_{ij}$. As a result, change in the total energy can result not just from a change in the spring elongation, but also from a change in the relative orientation of connected residues even if the distance between them is unchanged (see example on Figure 2.1). This is a direct consequence of the assumption that individual residue fluctuations are isotropic, i.e. they are identical in all directions. As such, directions become irrelevant, and only the amplitude of the individual fluctuations influences the potential energy.

Equation (2.14) can be conveniently rewritten as

$$V_{\mathrm{GNM}} = \frac{k}{2} \sum_{\alpha=1}^{3} \sum_{i<j}^{N} \left(q_i^\alpha - q_j^\alpha \right)^2$$

and, similarly to Equation (2.12) the Hessian matrix simply becomes

$$(\boldsymbol{H}_{\mathrm{GNM}})_{i,j} = \begin{cases} -k.I_3 & \text{if } i \neq j \text{ and } \|\boldsymbol{r}_{ij0}\| < r_c \\ \\ -\sum_{j \neq i} (\boldsymbol{H}_{\mathrm{GNM}})_{ij} & \text{if } i = j \end{cases} \tag{2.15}$$

[6] C_α atoms are the carbon atoms of the peptide bonds which link the amino acids in a chain. They are part of the backbone of the protein.

where I_3 is the 3×3 identity matrix, or equivalently,

$$\boldsymbol{H}_{\mathrm{GNM}} = k\boldsymbol{L} \otimes I_3 \tag{2.16}$$

where $\otimes$ designates the Kronecker product[7] and $\boldsymbol{L}$ the Laplacian matrix of the graph of the C_α atoms. Since all directions become equivalent and independent from each other, the connectivity between C_α atoms is the only property of the structure taken into account by the GNM and the protein can thus be conveniently represented as a graph.

Under the assumption that the residue fluctuations are Gaussian distributed, the residue fluctuation correlations can be shown to be given by the elements of the pseudo inverse $\tilde{\boldsymbol{L}}^{-1}$ of the Laplacian matrix (Kloczkowski *et al.*, 1989)

$$\left\langle \boldsymbol{q}_i^T \boldsymbol{q}_j \right\rangle = \frac{3}{\mathrm{k}} \left(\tilde{\boldsymbol{L}}^{-\mathbf{1}} \right)_{ij}$$

and the fluctuations in interresidue distances are therefore expressed by

$$\begin{aligned} \left\langle \boldsymbol{q}_{ij}^T \boldsymbol{q}_{ij} \right\rangle &= \left\langle \boldsymbol{q}_i^T \boldsymbol{q}_i \right\rangle + \left\langle \boldsymbol{q}_j^T \boldsymbol{q}_j \right\rangle - 2 \left\langle \boldsymbol{q}_i^T \boldsymbol{q}_j \right\rangle \\ &= \frac{3}{k} \left[\left(\tilde{\boldsymbol{L}}^{-\mathbf{1}} \right)_{ii} + \left(\tilde{\boldsymbol{L}}^{-\mathbf{1}} \right)_{jj} - 2 \left(\tilde{\boldsymbol{L}}^{-\mathbf{1}} \right)_{ij} \right]. \end{aligned}$$

Through the pseudoinverse of the Laplacian matrix, the GNM can thus compute the cross-correlations in residue fluctuations and inter-residue distances, and thereby reveal delocalised concerted motions, similarly to normal mode analysis.

Interestingly, this expression of interatomic distance fluctuations is also directly related to other concepts of graph theory such as the resistance distance in electrical networks (Klein and Randić, 1993), or the properties of a Markov process diffusing on the network of C_α atoms. In particular, considering a Markov process evolving on the graph, for instance a random walker jumping from node to node along the edges, the commute time of the process, i.e. the average time taken by the walker to go from node i to j and come back, is proportional to fluctuations in interresidue

[7] Using the Kronecker product, each element L_{ij} of the N×N matrix L is replaced by $L_{ij}I_3$, thus generating a 3N×3N matrix.

distances (Chennubhotla and Bahar, 2007)

$$\begin{aligned} C(i,j) &= \left[\left(\tilde{\boldsymbol{L}}^{-\mathbf{1}}\right)_{ii} + \left(\tilde{\boldsymbol{L}}^{-\mathbf{1}}\right)_{jj} - 2\left(\tilde{\boldsymbol{L}}^{-\mathbf{1}}\right)_{ij}\right] \sum_{l=1}^{N} d_l \\ &= \left\langle \boldsymbol{q}_{ij}^T \boldsymbol{q}_{ij} \right\rangle \left[\frac{k}{3} \sum_{l=1}^{N} d_l\right]. \end{aligned}$$

The GNM thus establishes a direct link between a Markov process evolving on the protein graph and residue fluctuations.

Hence, the Gaussian network model allows to calculate in an extremely efficient way fundamental properties such as individual residue fluctuations and, via the pseudo inverse of the Laplacian matrix, correlated motions. Its very low computational cost and the relatively good agreement with crystallographic B factors have made the GNM a very popular method to analyse protein structures. The GNM was found to reproduce X-ray and NMR data well (Yang *et al.*, 2007; Bahar *et al.*, 1997), and to broadly agree with molecular dynamics simulations although less so than the classical elastic network model coarse-grained at the residue level (Doruker *et al.*, 2000).

However, the Gaussian network model relies on a number of assumptions which must be carefully considered. This model was first introduced by James in 1947 (James and Guth, 1943; James, 1947) as an attempt to explain the elastic properties of rubber, following the work of Lord Rayleigh and Flory (Lord Rayleigh, 1919; Flory, 1969) on the statistics of ideal chain molecules. Rubber was then modelled as a network of individual freely jointed chains[8] interconnected at junction points. Between each pair of junction points, chains exert an entropic force that results from the decrease in the number of possible configurations of the chain when it is stretched, and which reproduces the force exerted by a spring. The isotropic assumption used by James was then motivated by the isotropicity of rubber. Forces in proteins are however not isotropic, have an enthalpic contribution and take place between atoms separated by much smaller distances.

The physical model underlying the GNM for proteins has consequently been criticised (Bahar *et al.*, 2007; Thorpe, 2007; Halle, 2002). In particular, the absence of rotational invariance in the GNM (i.e. rigid-body rotations yield an increase in potential energy, see Figure 2.1) has been pointed out by Thorpe (2007). This has consequences on the ability of the GNM to capture collective motions (Fuglebakk

[8]Simple model of polymer molecules which consists in a chain of connected rigid rods whose individual orientation is unconstrained by that of the neighbouring rods.

et al., 2013). For instance, the rigid-body rotation of a whole domain around a hinge, a common form of collective motion in proteins, would incur an unrealistically large energy cost in the GNM. Halle (2002) also observed that the profile of the amplitude of atomic fluctuations is to a large extent determined by the spatial variations in the local packing density[9] which, being well captured by a distance-based graph model, could explain the success of the GNM in reproducing the crystallographic B-factors (Rader *et al.*, 2006).

2.3 Rigidity analysis

Traditional engineering concepts of rigidity analysis have also been sucessfully applied to protein structures. Bonds and interactions of high energy define an ensemble of distance constraints which, to some extent, can be considered as rigid rods. Depending on their topology, they can rigidify some regions of a structure and leave other parts fully flexible. We briefly introduce two methods to identify rigid clusters in biomolecules, combinatorial rigidity and infinitesimal rigidity.

Combinatorial rigidity provides an algorithm to identify rigid clusters by a simple counting of the number of nodes and constraints. The method is based on Laman's theorem (Laman, 1970) which gives a necessary and sufficient condition to the existence of a rigid cluster by evaluating a simple inequality condition ($b \geq 2N - 3$) on the number of constraints b and the number of nodes N in all subgraphs[10] of the original structure.

As such, the direct application of the Laman's theorem scales badly with the number of nodes in the graph. However, it can be applied recursively using a particular algorithm, called the pebble game (Jacobs and Thorpe, 1995), which, in practice, scales linearly with the number of nodes. Unfortunately, Laman's theorem, originally formulated for 2D graphs, does not provide a sufficient condition anymore in the case of generic graphs in three dimensions. It is however directly applicable to a special kind of graph, coined "bond-bending networks", characterised by the presence of angle constraints between all next nearest neighbours (Jacobs, 1998) which limit the movable parts of the structure to the dihedral angles only.

The analysis of protein structures using combinatorial rigidity has been developed by Thorpe and coworkers in their software package FIRST (Jacobs *et al.*, 2001). Covalent bonds, salt bridges and hydrogen bonds are taken into account

[9] The number of atoms in a sphere of a small radius centered around the atom.
[10] All subsets of nodes and edges of the original graph.

directly as rigid rods together with two angle constraints while hydrophobic tethers are included by inserting an additional phantom node, which allows for sliding motions within the tethers.

Using a 3D version of the pebble game algorithm (Jacobs and Thorpe, 1995), redundant constraints and rigid clusters can be identified in most protein structures in a fraction of a second. Its linear scaling in both memory and cpu means that molecular structures of almost any size can be analysed by FIRST.

Other approaches for rigidity analysis of proteins have also been proposed. In particular, Costa (2008) introduced the use of infinistesimal rigidity to extracts the subspace of infinitesimal motions which respect the ensemble of N_c fixed distance constraints c_{ij}

$$\left\| \boldsymbol{r}_i - \boldsymbol{r}_j \right\|^2 = c_{ij}^2 \tag{2.17}$$

for the N_c pairs of atoms (i, j) subject to a constraint based on the same ensemble of covalent bonds and weak interactions as FIRST. Deriving Equation 2.17 yields an ensemble of conditions which must be respected by infinitesimal displacements $\dot{\boldsymbol{r}}_i = \frac{\mathrm{d}\boldsymbol{r}_i}{\mathrm{d}t}$ of the constrained atoms

$$(\boldsymbol{r}_{i0} - \boldsymbol{r}_{j0})\dot{\boldsymbol{r}}_i - (\boldsymbol{r}_{i0} - \boldsymbol{r}_{j0})\dot{\boldsymbol{r}}_j = 0 \tag{2.18}$$

or, in matrix form,

$$C\dot{\boldsymbol{r}} = 0$$

where the matrix C is the $\mathrm{N}_c \times 3\mathrm{N}$ rigidity matrix and $\dot{\boldsymbol{r}}$ the $3\mathrm{N} \times 1$ vector of infinitesimal displacements. Hence, the null space of the rigidity matrix C defines the space of infinitesimal motions which respects all N_c constraints. Rigid clusters can then be identified by monitoring the atoms that keep their relative orientation unchanged after infinitesimal displacements.

Infinitesimal rigidity is applicable to any 3D constraint framework, without being limited to bond-bending networks only. It comes however at the expense of a higher computational cost, scaling as $\mathcal{O}(N^3)$ in the worst cases and $\mathcal{O}(N^2)$ when all constraints are independent. A heuristic has also been proposed (Costa, 2008) which scales linearly but only guarantees a lower bound on the number of degrees of freedom.

The mechanical point of view of rigidity theory provides an intuitive perspective on the structural organisation of biomolecules. In addition, the presence of hinges, rigid clusters, and domains displaying concerted motions, which can all be rapidly

identified through rigidity analysis, are often key to the function of most protein and DNA molecules. Considering the ability of weak interactions to break and reform to varying degrees depending on their energy, the scope of these methods can be extented by studying changes in the rigid clusters as the energy threshold in the graph construction changes. Finally, beyond the insight they give on the structural organisation of the protein, they also provide an excellent coarse-graining for simulations.

Chapter 3

Methodology

Most of the computational methods described in the previous chapter were designed to probe one particular time or spatial scale, whether it be slow modes involving concerted motion over the whole structure or localized rigid clusters. However, none really provides a way to seamlessly link the atomic, residue, secondary, tertiary and quaterary levels of organisation, or is able to trace back the emergence of the large scale, slow mode, behaviour from the structural organisation of the biomolecule at the atomic level.

The objective of this work is to uncover dynamical and functional properties of proteins, not only at the domain level, but over the entire range of scales, from atoms to the quaternary structure and beyond. This is enabled by the complete characterisation of the physico-chemical details of the biomolecular structure into a network of interconnected atoms, combined with a dynamical multi-resolution graph partitioning framework called Markov stability. We now describe in details the motivation and principles of the methodology at the basis of this work.

3.1 Motivation

The success of the methods introduced in the previous chapter provides clear evidence of the considerable insight that can be gained about the function and dynamics of proteins from the analysis of the static structure alone. However, as is often the case in complex systems, the intrinsic structural organisation of large biomolecules is concealed by an extensive, intricate and diverse ensemble of interactions taking place between the many atoms of which they are constituted. An immediate consequence

of this complexity is an often perceived trade-off between the level of description that should be used by a method and the time and spatial scales it can access.

The ensemble of interatomic bonds and interactions in large biomolecular structures however naturally lends itself to a graph theoretical description. In addition to being intuitive and appropriate, rationalising a protein structure as a network of interacting atoms unlocks a whole range of graph theoretical tools which can unravel, at a very low computational cost, key properties of the network which relate to features of biological interest of the protein structure (see Section 2.2.1 and reviews by Csermely *et al.* (2013); Di Paola *et al.* (2013); Böde *et al.* (2007)).

Here, we take advantage of the low computational cost associated with graph theoretical methods to probe over the entire range of scales the structural organisation that defines the dynamics. Importantly, this analysis is carried out from the atomic description, and without the use of any a priori information other than the topology of interatomic connections specified by the physico-chemical bonds and interactions of the original structure. To this end, we make use of a multi-resolution graph partitioning method. Tools for the identification of community structures have already been shown to be powerful methods for the analysis of protein structures, whether it be to uncover allosteric mechanisms, study protein stability, or establish the functional domains (Kannan and Vishveshwara, 1999; Chennubhotla and Bahar, 2006; del Sol *et al.*, 2007; Sethi *et al.*, 2009; Kundu *et al.*, 2004). Remarkably, community structures have also been used in the study of intramolecular signals with residues at the interface between communities suggested as essential to the communication pathways (del Sol *et al.*, 2007; Sethi *et al.*, 2009; Chennubhotla and Bahar, 2006). Here, rather than studying the impact of the community structure on the communication accross the molecule, we use the propagation of a Markov process on the network of atoms over different time scales as a way to explore the structural organisation of the protein and relate it to its biological function.

3.2 Contributions and summary of previous work

The generic methodology introduced in this chapter was first suggested as a successful route for the analysis of the multiscale organisation of protein structures by Delvenne *et al.* (2010) and subsequently refined by Meliga (2009).

The structure was originally modelled by Delvenne *et al.* (2010) as an unweighted network of covalent bonds, hydrogen bonds, salt bridges and hydrophobic tethers at the atomic level. Subsequently, the method has been revised by Meliga (2009) using

an approximated spring constant derived from the energy potentials. In this work, the graph theoretical model was further improved into a fully consistent energy-based graph. In particular, strong electrostatic interactions and π-stacking interactions have been included. It has also been also further extended to nucleic acid structures such that any biomolecule, whether it be DNA or proteins, can be analysed and non-standard amino-acids or ligands can be included.

The mathematical developments of Markov stability presented below were first introduced by Delvenne *et al.* (2010) and the random walk interpretation was proposed by Lambiotte *et al.* (2009). The latter also introduced the use of the variation of information to identify the most relevant levels of organisation in the all-scale analysis given by Markov stability. The contribution of this work lies here in bringing all these different components together into a general framework for the identification of the biochemically meaningful substructures in biomolecules. In the following chapters, the methods is further expanded through the use of random graph surrogates, the analysis of the landscape of optimised partitions, and the detection of structurally important edges in the graph (mutational analysis). A Matlab/C++ code was also developed as part of this work for the use of the Louvain algorithm for Markov stability.

3.3 Modelling biomolecules as networks

A variety of strategies have been proposed over the recent years to construct a graph from the spatial coordinates of a biomolecular structure (reviewed by Di Paola *et al.* (2013); Csermely *et al.* (2013); Böde *et al.* (2007)). The vast majority relies on an amino acid level of description, each node corresponding to one residue, with edges, often unweighted, defined by inter-residue contact maps using a euclidian distance cut-off between all pairs of residues. Other methods have also been proposed with edges weighted based on the number of atom-atom contacts (Kannan and Vishveshwara, 1999; Chennubhotla and Bahar, 2006) or correlation coefficient computed from short molecular dynamics trajectories (Sethi *et al.*, 2009).

The rationale behind the construction of the graph is however rarely discussed and distance-based edge assignments is often assumed as a default choice with highly variable distance cut-offs ranging from 3 to 18Å(Soheilifard *et al.*, 2008; Park and Kim, 2011; Csermely *et al.*, 2013). Yet all the graph theoretical properties derived from the protein network, whether it be community structures, shortest paths, centrality, clustering coefficient or degree distribution are highly dependent on the cri-

teria chosen by the modeller to assign edges and edge weights. Schemes based on distance cut-off can be approriate in certain situations. For instance, as detailed in the previous chapter, Tirion (1996) demonstrated that a simple distance cut-off is sufficient to reproduce the slow modes in elastic network models. This property was also later successfully reused by Bahar *et al.* (1997) in the protein graph underlying the Gaussian network model. While the success of elastic network models in reproducing large conformational transitions supports the idea that slow modes are insensitive to the details of the interatomic potentials, it is unclear whether allosteric mechanisms and signal propagation in the protein should be equally robust to the way edges are assigned, especially as some graph theoretical measures, such as shortest paths, dramatically rely on the exact location of the edges. Bongini *et al.* (2010) notably suggested that outside of the low end of the frequency spectrum, key vibrational properties of the secondary structure can only be reproduced when differences in the bond strength are taken into account in the network.

The challenge is therefore to find a simple yet physically realistic graph model which retains the key biophysical properties involved in the phenomena studied. When using distance-based edge assignments, one ignores the physico-chemical details of the interactions and only the relative local spatial position of the residues is captured in the graph. In many graph construction schemes, covalent interactions are even explicitely removed (Ribeiro and Ortiz, 2014; Vijayabaskar and Vishveshwara, 2010), despite the fact that the motions of covalently bound atoms are usually highly correlated (Ichiye and Karplus, 1991). Yet it seems reasonable to expect signals that propagate structural changes to travel from one residue to another along the bonds and interactions that exist between them. Through a comparison with experimental and computational studies, recent research (Ribeiro and Ortiz, 2014) showed that edge weights computed from the inter-residue interaction energy (Vijayabaskar and Vishveshwara, 2010; Jiao *et al.*, 2007) lead to a more accurate reproduction of the intra-protein signaling pathways than graphs constructed from distance cut-offs or correlated motions.

As we here wish to exploit the diffusion of signals on the graph to probe the structural organisation of biomolecules all scales, the topology of distance-based inter-residue contacts is unlikely to be sufficiently accurate and the physico-chemical details of the interactions should be fully taken into account. In this work, biomolecules are therefore encoded in terms of a weighted graph formalism that is built from the atomistic description of the structure using the potential energy derived from atomic force fields. In this formalism, edge assignment and weighting is thus entirely based

on the underlying chemistry, using criteria based on geometry, interaction energy and atom type. Following from observations of Ribeiro and Ortiz (2014), edges are weighted by the interaction energy, and signals are thus assumed to travel preferentially along higher energy bonds and interactions. Finally, the graph is defined at the *atomic* level. Bonds and interactions indeed fundamentally take place between atoms rather than between residues. Amino acids also exhibit strong variations in size and physico-chemical properties, with triptophan containing more than three times the number of atoms of glycine. In addition, coarse-graining the network at the residue level is unnecessary considering the low computational cost of graph theoretical methods. Hence, rather than being preimposed in the construction of the network, we expect the individual residues to appear naturally as a result of our analysis.

3.3.1 Edge assignment and weighting

Our methodology builds upon the work of Jacobs *et al.* (2001) on the combinatorial rigidity analysis of protein structures, which defines a sparse graph at the atomic level with a detailed description of the covalent bonds and weak interactions. The edge weighting and assignment methodology presented in this section is a revised and improved version of those originally proposed by Meliga (2009) and Delvenne *et al.* (2010). In particular, it has been adapted to the analysis of DNA molecules, is energy-based, and includes electrostatic as well as π-stacking interactions.

The process of the construction of the graph is summarised in Figure 3.1. The graph is generated from the atomic spatial coordinates of an all-atom experimental structure (X-ray or NMR) obtained from the PDB database (Bernstein *et al.*, 1978). For X-ray crystal structures, missing hydrogen atoms are first added using the software package Reduce (Word *et al.*, 1999). Each atom in the experimental structure is then included as a node in the graph, and all covalent bonds and weak interactions (hydrogen bonds, salt bridges, hydrophobic tethers, large electrostatic interactions and π-stacking interactions) between a pair of atoms are represented by an energy-weighted edge in the graph derived from an atomic force field (summarized in Table 3.1). Full details of the procedure, potential function, and parameters used can be found in Appendix A. We now briefly summarize the assignment and weighting of edges in the graph.

Edges for the covalent bonds, hydrogen bonds, salt bridges and hydrophobic tethers are all identified in the structure from the geometric criteria implemented in

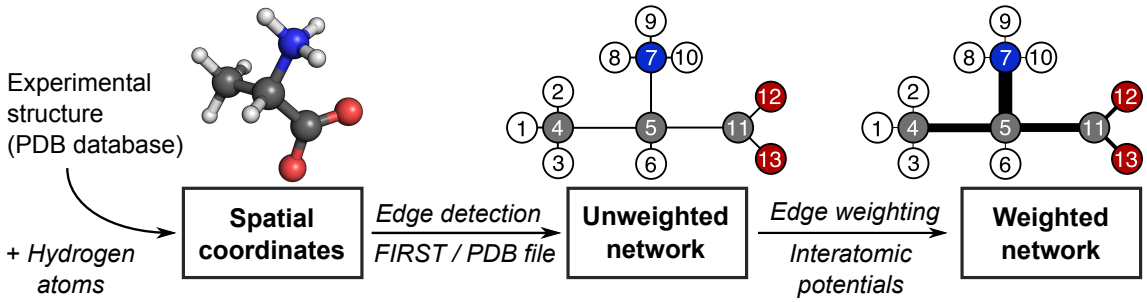


Figure 3.1: Construction of the weighted graph from an experimental structure.

the software package FIRST (Thorpe, 2009; Jacobs *et al.*, 2001). Each edge is then given a weight equal to the energy of interaction between the two atoms linked and derived from specific potential associated with each type of interaction. The energy of covalent bonds is obtained from standard tabulated values of bond dissociation energies (Huheey *et al.*, 1993). Hydrophobic tether edges are weighted using the hydrophobic potential of mean force proposed by Lin *et al.* (2007), and hydrogen bonds and salt bridges using the modified Mayo potential proposed by Rader *et al* (Rader *et al.*, 2002; Dahiyat *et al.*, 1997).

Π stacking interactions are modelled differently in proteins and DNA. In proteins, they are identified using FIRST and assigned a fixed weight of 10 kcal/mol corresponding to a typical energy for this type of weak interaction (Sponer *et al.*, 2008). In DNA however, stacking interactions are much more abundant and play a major role in the dynamics of nucleic acid molecules. We thus chose to model them more accurately using the DNA-specific potential proposed by Hunter and Sanders (1990), which sums the contributions from van der Waals and electrostatic interactions. In DNA, the stacking interaction assignments given by FIRST using the relative orientation of bases is ignored and the edges are assigned based on an energy threshold on the Hunter & Sanders potential.

Due to their long-range nature, we made the choice to generally neglect electrostatic interactions in the graph. In addition to being small, they decay very slowly with the distance and including them would thus lead to an almost fully connected graph which would simply blur the details of the dominant paths of communication. Strong electrostatic interactions which were known to play a crucial role in the dynamics or function of the protein or DNA, i.e. coordination of metal ions and electrostatic interactions between phosphate groups of the DNA backbone, are however included in the graph and weighted using the Coulomb potential and point

Interaction type	Potential
Hydrogen bonds & salt bridges	Modified Mayo potential (FIRST) (Jacobs *et al.*, 2001; Dahiyat *et al.*, 1997)
Hydrophobic tethers	Hydrophobic potential of mean force (Lin *et al.*, 2007)
π-stacking interaction	Sum of Van der Waals and electrostatic interactions (Hunter and Sanders, 1990)
Electrostatic interactions	Coulomb potential using partial charges from the all-atom OPLS-AA force field (Jorgensen and Tirado-Rives, 1988)
Covalent bonds	Tabulated dissociation energies (Huheey *et al.*, 1993)

Table 3.1: Force field used for weighting the edges of the graph of a protein structure.

charges derived from the OPLS-AA molecular dynamics force fields (Jorgensen and Tirado-Rives, 1988).

This framework is however only one of the many ways in which the graph can be constructed. Different force fields and physical criteria can be used for edge assignment and weightings which would similarly lead to a biologically relevant representation of the biomolecule. Although the use of a weighted graph is crucial for our analysis, our numerics show that our final results are relatively insensitive to small variations in the edge weights or the atomic resolution (see Appendix B). The framework introduced here and used throughout this work is thus robust with respect to the details of the potentials, and energy-weighted networks constructed from any atomic force field should yield essentially identical results.

3.4 Community structures in networks

Due to the large number of elements combined with their non-trivial relationships, it is often hard to understand the global behaviour of a large network by focusing solely on its individual constituents. When dealing with complex graphs, it is therefore sometimes desirable to obtain simplified reduced representations in terms of subgraphs or *communities*, i.e. meaningful groupings of nodes that are significantly related. This coarse-grained representation provides a mesoscopic scale perspective on its overall organisation with the objective to gain a deeper understanding of its function and general properties.

Although intuitive, community structure in networks still lacks a rigorous definition, and a large number of methods have been introduced over the recent years, traditionally expressing the quality of a graph partition[1], in terms of the density of edges. One commonly accepted notion of a community is that of a tightly-knit group with many connections within the group and fewer to external nodes.

This form of modular architecture is commonly found in many networks spanning various fields including biology, engineering and sociology. Well-known examples include social communities or online social networks, but also neural networks, the Internet, the world wide web and, the subject of this work, large biomolecules. Community detection has a long history, and recent research following the pioneering work of Newman and Girvan (2004), has both rediscovered classic results and introduced novel methods (Fortunato, 2010; Schaeffer, 2007).

3.5 The Markov stability of a graph partition

The structural organisation of biomolecules such as proteins is however not only complex due to the large number of interactions, it is also defined over multiple scales, as discussed in Chapter 1. While many graph partitioning methods aim at identifying the single best division of the graph into communities, we are looking to explore the structural organisation of biomolecules at *all scales*.

Beyond the structure-based approaches which use the location of edges, community detection can also be approached from a *dynamical* perspective. Indeed, the temporal evolution of a dynamical process is tightly linked to the topology of the network on which it unfolds, and the observation of one can often reveal key characteristics of the other (Strogatz, 2001; Barahona and Pecora, 2002). Accordingly, several methods have taken this approach to unveil communities, using processes such as random walks (Gfeller and De Los Rios, 2007; van Dongen, 2000) or the synchronisation of oscillators (Arenas *et al.*, 2006).

Interestingly, Delvenne *et al.* (2010) showed that the dynamical viewpoint provides a route towards elucidating the underlying community structure in a network over the entire spectrum of scales. This method called *the Markov stability of a graph partition* (Delvenne *et al.*, 2010; Lambiotte *et al.*, 2009; Delvenne *et al.*, 2012) allows to establish the optimal community structures at all levels, from communities including a single element, to the optimal division of the whole system into two

[1] A graph partition is here defined as the subdivision of the nodes of a network into an ensemble of non-overlapping communities.

groups, by using the time evolution of a Markov process on the network. Communities are here seen as regions which temporarily trap the dynamical process evolving on the network. Considering a random walker jumping from node to node along the edges of the graph, the walker would be expected to remain with a high likelyhood within a group of densely connected nodes, since links to other nodes of this community are more frequently encountered by the walker in this region than towards other parts of the graph. As the dynamical process is allowed to explore the graph for longer, it becomes more likely to escape smaller communities and gets trapped over larger portions of the network. More precisely, a partitioning will be associated with a high Markov stability over a particular time scale if the dynamical process tends to be more contained inside the communities over that time scale than would otherwise be expected at stationarity (i.e. after an infinite time).

Formally, we consider the general case of a random walk taking place on the graph whose dynamics is driven by a continuous time Markovian diffusion process[2]

$$\dot{\boldsymbol{p}}(t) = \boldsymbol{p}(t)\boldsymbol{Q}. \tag{3.2}$$

Here, $\boldsymbol{p}(t)$ denotes the 1×N vector of probabilities representing the density of random walkers on each of the N nodes of the graph at time t. The N×N matrix $\boldsymbol{Q}$ encodes the dynamics that governs the time evolution of the random walk, where Q_{ij} defines the rate at which the random walker jumps from node i to node j. As such, we effectively define a poisson process on each node of the graph with an exponentially distributed waiting time with mean $1/Q_{ij}$. Since the random walkers progress along the edges, $\boldsymbol{Q}$ is directly defined by the topology of the graph and is typically contructed from its N×N adjacency matrix.

The solution from Equation 3.2 is given by

$$\boldsymbol{p}(t) = \boldsymbol{p}(0)e^{\boldsymbol{Q}t}, \tag{3.3}$$

[2]Although we will use this continuous time process throughout the rest of this work, it is worth noting that discrete time processes of the form

$$\boldsymbol{p}_{t+1} = \boldsymbol{p}_t\boldsymbol{Q}. \tag{3.1}$$

can also be used. This yields a different version of Markov stability associated with synchronous jumps taking place at unit time intervals, as opposed to jumps occuring at random times in the continuous case. Further discussion on discrete time Markov stability can be found in references (Delvenne *et al.*, 2010, 2012).

and defines, for each node, the probability that the random walker will be at the same position initially and at time t (having potentially visited other nodes in the time interval).

We now use the time evolution of this random process as a way to probe the underlying structure of the graph. As we have indicated, communities can be viewed as subgraphs with a superior ability to retain the probability flow. The autocovariance matrix of the random process provides a natural way to reveal regions where the random walk is transiently trapped by measuring the similarity of the process with itself at a later time. The autocovariance of the random process can be expressed as

$$\boldsymbol{X}(t) = \mathrm{cov}(\boldsymbol{p}(0), \boldsymbol{p}(t)) = \mathrm{E}[\boldsymbol{p}(0)^T \boldsymbol{p}(t)] - \mathrm{E}[\boldsymbol{p}(t)]^2 = \boldsymbol{\Pi} e^{\boldsymbol{Q}t} - \boldsymbol{\pi}^T \boldsymbol{\pi}, \tag{3.4}$$

where E is here the expectation, and we here made use of the ergodicity of the stochastic process, with $\boldsymbol{\pi}$ designating its stationary distribution. $\boldsymbol{\Pi} = \mathrm{diag}\,(\boldsymbol{\pi})$ is the diagonal matrix of the stationary distribution, taken here as the initial condition of the process, which is a reasonable assumption in the absence of any *a priori* information.

The entry $X_{ij}(t)$ effectively measures the fraction of the total probability flow that has been transferred from node i to node j after a time t, weighted by the initial probability at node i at time zero, discounted by the expected probability flow that would be transfered at random. To evaluate the quality of a particular (hard) partition $\mathcal{P}$ in terms of the total probability flow retained within each of its M communities, we need to compute the same matrix $\boldsymbol{X}$, but now in terms of the *clustered* random walk taking place at the community level. To this end, we define the N×M indicator matrix $\boldsymbol{H}$ of $\mathcal{P}$ with entries H_{ij} equal to one if node i belongs to community j and zero otherwise. Using the linearity of the covariance matrix, the autocovariance of the *clustered* random walk $\boldsymbol{y(t)} = \boldsymbol{H}^T \boldsymbol{p(t)}$ is given by the M×M matrix

$$\boldsymbol{R}(t, \mathcal{P}) = \boldsymbol{H}^T \left[\boldsymbol{\Pi} e^{\boldsymbol{Q}t} - \boldsymbol{\pi}^T \boldsymbol{\pi} \right] \boldsymbol{H}. \tag{3.5}$$

The effect of the matrix $\boldsymbol{H}$ is to sum all the entries of $\boldsymbol{X}$ corresponding to the nodes associated with each community of the partition, and $(\boldsymbol{R}(t, \mathcal{P}))_{ij}$ thus measures the excess probability of a random walker which started in community i to end up in community j at time t over the expectation of it happening by chance.

Since we are here only interested in the events where the random walkers remain in the same community, whose probabilities are given for each walker by the diagonal

elements of $\boldsymbol{R}(t, \mathcal{P})$, we define the *Markov stability of a graph partition* $\mathcal{P}$ as

$$r(t, \mathcal{P}) = \text{trace}\left(\boldsymbol{R}(t, \mathcal{P})\right). \tag{3.6}$$

From the random walk perspective, Markov stability can thus equivalently be expressed in terms of the probability of the random walker to be found in the same community after a period of time t

$$r(t, \mathcal{P}) = \sum_{\mathcal{C} \in \mathcal{P}} P_{\mathcal{C}}(0, t) - P_{\mathcal{C}}(0, \infty), \tag{3.7}$$

where $P_{\mathcal{C}}(0, t)$ denotes the probability of the random walker to be in community $\mathcal{C}$ at times 0 and t.

The dynamics $\boldsymbol{Q}$ can be defined in a number of different ways, which provides the opportunity to include any *a priori* information one may have about the problem at hand and the way nodes communicate or interact with each other in the system analysed. In particular, the dynamics of many physical systems, such as electrical networks (Schaub *et al.*, 2014; Wu and Huberman, 2004) and oscillators (Barahona and Pecora, 2002; Arenas *et al.*, 2006), is governed by the Laplacian matrix $\boldsymbol{L}$ of the graph [3], in which case $\boldsymbol{Q} = \boldsymbol{A} - \boldsymbol{D} = -\boldsymbol{L}$, where $\boldsymbol{A}$ designates the adjacency matrix of the graph and $\boldsymbol{D}$ the diagonal matrix of the nodes degree $D_{ii} = k_i = \sum_j A_{ij}$. Unsurprisingly, this dynamics is one of the most commonly used. In the particular case of proteins, the dynamics driven by the Laplacian matrix has been successfully used to model signal propagation (Chennubhotla and Bahar, 2007, 2006) and diffusion of vibrational dynamics (Reuveni *et al.*, 2010a) throughout biomolecular structures, and this will also be the dynamics used throughout this work.

Using the Laplacian matrix, the stationary distribution becomes $\boldsymbol{\pi} = \mathbf{1}_{\boldsymbol{N}}/N$, where $\mathbf{1}_{\boldsymbol{N}}$ the 1×N vector of ones, and is uniform over all the nodes in the network. Markov stability can be rewritten as

$$r(t, \mathcal{P}) = \text{trace}\left(\boldsymbol{H}^T \left[\frac{1}{N} e^{-\boldsymbol{L}t/\langle k \rangle} - \frac{\mathbf{1}^T\mathbf{1}}{N^2}\right] \boldsymbol{H}\right), \tag{3.8}$$

where $\langle k \rangle$ is a normalisation factor equal to the average degree of the graph.

For very large graphs (typically $N > 10000$), evaluating the exponential in Equation 3.5 can be computationally very demanding. When the computational resources

[3] More precisely, the standard Laplacian, also called combinatorial Laplacian.

do not allow to compute the full Markov stabiliy $r(t, \mathcal{P})$, we use a linearisation given by the first order expansion of $r(t, \mathcal{P})$ around zero

$$r_{lin}(t, \mathcal{P}) = r(0, \mathcal{P}) + t \left. \frac{\mathrm{d}r(t,\mathcal{P})}{\mathrm{d}t} \right|_{t=0} = r(0, \mathcal{P}) - \mathrm{trace}\left(\boldsymbol{H}^T \left[\frac{\boldsymbol{L}}{2m} \right] \boldsymbol{H} \right). \tag{3.9}$$

where $2m = \sum_i k_i$ is the sum of the degrees of the nodes in the graph, or two times the sum of all the edge weights.

Hence, Markov stability is defined for a particular *Markov time* t associated with the dynamics that reveals the community structure. The analysis can be viewed as following the time evolution of a probabilistic process on the graph and identifying the subgraphs where the probabilistic flow gets trapped. As the Markov time increases, Markov stability follows the expanding transient of this dynamics towards stationarity and, in doing so, it allows us to reveal naturally a sequence of coarser partitions that uncovers the multiscale structure of the graph, if it exists.

Unlike traditional partitioning methods, Markov stability is thus not limited by an implicit scale. The concept of flow at its base allows for instance to detect non-cliquish communities, i.e. communities that are not characterised by the local node-level density of links but by the broader view of retention of flow (Schaub *et al.*, 2012). The Markov time here acts as zooming lens allowing to focus the analysis on a chosen scale without any upper or lower limit. By sweeping the entire range of Markov times, Markov stability decomposes the structure into a hierarchy of communities which reflects its levels of organisation at all scales. This property makes it particularly well suited for the analysis of protein structures, which possess an intrinsic organization spanning a vast range of scales (from chemical groups to functional domains) with non-clique like communities which are difficult to detect using standard methods (Schaub *et al.*, 2012).

Markov stability also provides a dynamical interpretation of community structures that generalises classical heuristics for graph partitioning. In particular, the partitioning according to the Fiedler eigenvector is equal to Markov stability for large times going to infinity (Delvenne *et al.*, 2010; Lambiotte *et al.*, 2009; Delvenne *et al.*, 2012) and Markov stability at time zero can be linked to measures of entropy. Finally, when using a discrete Markov chain or the linearised version of the con-

tinuous time Markov stability, the very popular modularity (Girvan and Newman, 2002) is identical to Markov stability at Markov time one.[4]

3.6 The Louvain algorithm

Markov stability, as defined in equation (3.6), provides a measure to assess the quality of a defined partition. However, as it is the case in most clustering-related problems (Fortunato, 2010), the global optimisation of Markov stability is computationally hard (Delvenne *et al.*, 2010; Brandes *et al.*, 2008)—a common occurrence in the study of complex landscapes. Finding the exact solution for the optimal Markov stability partition is thus impossible for all but the smallest graphs. In practice, a variety of heuristic strategies can however be used to obtain good partitions which can then be ranked by Markov stability to provide us with near-optimal partitions at different time scales. Several such algorithms exist, and often proceed by either progressively aggregating nodes (agglomerative algorithm) or gradually dividing the whole network into smaller and smaller groups of nodes (divisive algorithm).

Here, we use a greedy agglomerative method, the Louvain algorithm (Blondel *et al.*, 2008), which has been shown to provide an extremely efficient optimisation of Markov stability. Briefly, Louvain works as follows. Initially, each node is assigned to its own community, i.e. the number of communities equals the number of nodes. Then, each node is transferred in turn into the neighbouring community (i.e. a community to which it is linked by an edge) where the increase of Markov stability is the biggest, as long as it improves the Markov stability of the overall partition. This step is repeated until no further transfer can increase the Markov stability. At that point, a new meta-graph of communities is generated, and the algorithm repeats these two steps until a coarse-grained graph is obtained where no further grouping can improve the Markov stability.

The Louvain algorithm has been observed to require little computational effort and to find partitions close to the optimal solution (Blondel *et al.*, 2008). The method is deterministic but the final solution found depends on the order in which the different nodes are scanned for the grouping step. This initial ordering, which we will refer to as the *Louvain initial condition*, can be chosen at random every time the algorithm is run. Indeed, we will use the variability of the observed solution induced

[4]Please note that, to allow for comparison between the Markov stability analysis of different protein structures, the Markov time is normalised by the total number of nodes throughout this work ($t_{\text{norm}} = t.N$).

by our random choice of the Louvain initial condition to estimate the robustness of a partition, a measure of its relevance.

3.7 The variation of information

For each value of the Markov time, a different partition with optimal Markov stability can be obtained. However, there should only be a limited number, if any, of meaningful levels of organisation in any network. Consequently, not all scales should be relevant, and not all Markov times should lead to a significant community structure. As is the case for modularity (Good *et al.*, 2010; Karrer *et al.*, 2008), the Markov stability value alone is a poor predictor for the relevance of a graph partition and an independent measure is needed to discriminate meaningful partitions from transient solutions. This issue is here addressed using a robustness tool, which is of general importance in multiscale analysis methods.

As suggested by Karrer *et al.* (2008), a distinctive property of a significant community structure should be its robustness to small perturbations. The expectation is that good partitions should be clear and well-defined in the network, and therefore relatively insensitive to noise. Upon introducing slight variations in the graph itself, the partitioning heuristic, or quality function, the new partition found should be highly similar to the one obtained originally. In this sense, the "Markov lifetime" of a partition, i.e. the Markov time span for which the partition is optimal in terms of Markov stability, provides a straightforward way to obtain an initial assessment of its robustness and relevance, the perturbation being here given by the change of Markov time itself (Ronhovde and Nussinov, 2009; Meliga, 2009; Delvenne *et al.*, 2010).

An alternative way consists in quantifying the extent to which the partitions are affected by the perturbation (Ronhovde and Nussinov, 2009; Karrer *et al.*, 2008; Good *et al.*, 2010) using a measure of distance between the solutions found before and after the perturbation. An information-theoretic distance between two partitions can be measured by the *variation of information* (Meila, 2003, 2007, 2005), a true metric based on the total information which is not shared by two partitions.

Consider a partition $\mathcal{P}$ of a graph into M communities C_k, $k = 1...M$ of n_k nodes. If we choose a node at random, how much uncertainty is there about the community it is assigned to in $\mathcal{P}$? The probability of this node belonging to community k can be estimated by the fraction of the nodes in C_k, $P(k) = \frac{n_k}{N}$. From these probabilities, one can use the common measure of uncertainty given by the Shannon entropy,

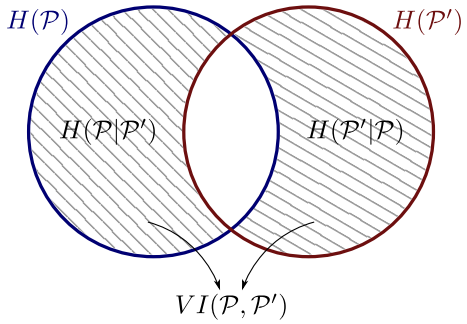


Figure 3.2: The variation of information (hatched region) is the sum of $H(\mathcal{P}|\mathcal{P}')$ the additional information needed to describe $\mathcal{P}$ given $\mathcal{P}'$ and $H(\mathcal{P}'|\mathcal{P})$ the additional information needed to describe $\mathcal{P}'$ given $\mathcal{P}$

which, for nodes allocations in partition $\mathcal{P}$ can be defined as

$$H(\mathcal{P}) = -\sum_{k} P(k) \log P(k). \tag{3.10}$$

The variation of information estimates the distance between two partitions $\mathcal{P}$ and $\mathcal{P}'$ as the sum of the uncertainty left about $\mathcal{P}$ when knowing $\mathcal{P}'$ and the uncertaintly left about $\mathcal{P}'$ when knowing $\mathcal{P}$, which can be expressed as

$$VI(\mathcal{P},\mathcal{P}') = H(\mathcal{P}|\mathcal{P}') + H(\mathcal{P}'|\mathcal{P}) \tag{3.11}$$

where $H(\mathcal{P}|\mathcal{P}')$ designates the entropy of $\mathcal{P}$ conditional on $\mathcal{P}'$, or equivalently

$$VI(\mathcal{P},\mathcal{P}') = 2H(\mathcal{P},\mathcal{P}') - H(\mathcal{P}) - H(\mathcal{P}'), \tag{3.12}$$

where $H(\mathcal{P},\mathcal{P}') = -\sum_k P(k,k') \log P(k,k')$ designates the joint entropy of $\mathcal{P}$ and $\mathcal{P}'$, with $P(k,k') = \frac{n_{k,k'}}{N}$ the probability that a node belongs to community k in $\mathcal{P}$ and k' in $\mathcal{P}'$. In this sense, the variation of information measures the total information content which is not in common between $\mathcal{P}$ and $\mathcal{P}'$.

As such, the variation of information depends on the size of the network: larger networks contain more information. The maximum variation of information achievable between two partition is given by $\log N$ and, to allow for comparisons, we will thus use the normalized VI (Meila, 2007) $VI_{\text{norm}}(\mathcal{P},\mathcal{P}') = VI(\mathcal{P},\mathcal{P}')/\log N$ in the rest of this work.

A variety of other measures of distance between partitions exist, often based either on counting pairs of nodes assigned in the same and different clusters in both partitions, such as the Rand index (Rand, 1971) and Jaccard coefficient (Ben-Hur

et al., 2002), or on matching communities between both partitions (for a detailed review and comparison of the different methods with the VI, see reference (Meila, 2007)). However, the Variation of Information posesses several desirable and intuitive properties, such as being a true metric on the space of partitions (i.e. it respects the properties of a distance measure such as the triangle inequality), which are not fully met by other methods (Meila, 2005).

The random initial conditions of the Louvain optimisation algorithm provide us with an ideal perturbation with respect to which we can measure the robustness of the partitions. By optimizing Markov stability for an ensemble of such initial conditions for each Markov time, we can calculate the variation of information between all pairs of solutions obtained, and compute the average as a measure of the relevance of the solutions obtained at a particular scale. Other perturbations affecting for instance edge weights or the quality function itself have been considered in the past and shown to yield similar results (Lambiotte, 2010).

Chapter 4

Markov stability analysis of adenylate kinase

In this chapter, we exemplify the Markov stability, variation of information developed in the previous chapter on adenylate kinase (AdK) from *Escherichia coli* and introduce two biochemically motivated surrogate random graph models. AdK is a classic example of a protein whose structure and dynamics have been studied extensively both experimentally and computationally (Henzler-Wildman *et al.*, 2007a), and thus serves as an ideal case study to illustrate and evaluate the capabilities of the computational framework we just introduced.

4.1 AdK structure and function

Adenylate kinase (AdK) is a small monomeric phosphotransferase enzyme that catalyses the reversible transfer of a phosphoryl group from adenosine triphosphate (ATP) to adenosine monophosphate (AMP) to yield two molecules of adenosine diphosphate (ADP) via the reaction:

$$\text{ATP} + \text{AMP} \overset{\text{Mg}^{++}}{\rightleftharpoons} 2\ \text{ADP}.$$

The structure of AdK is characterised by three domains: the LID, where ATP binds, CORE, and AMP binding domains (see Figure 4.1). Structural analyses (Müller *et al.*, 1996; Müller and Schulz, 1992) showed that AdK takes an open conformation when unliganded and a closed form when bound to the substrates or some inhibitors. The transition between the two forms involves the LID and the AMP domains closing over the CORE region, which provides the condition for catalysis by optimally positioning the AMP and ATP molecules and shielding the active site from the solvent (Maragakis and Karplus, 2005; Gerstein *et al.*, 1993; Müller *et al.*,

1996). Although the closed conformation is the catalytically active form in which the phosphorylation takes place, AdK has been observed to spontaneously explore the closed conformation even in the absence of substrate (Arora and Brooks III, 2007; Henzler-Wildman *et al.*, 2007b).

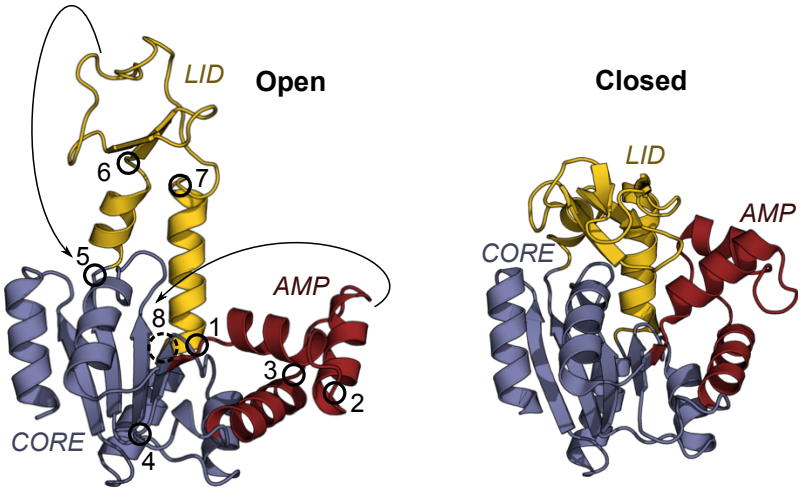


Figure 4.1: The conformational change between the free (left, PDB 4AKE) and the complex-bound AdK (right, PDB 1AKE) is characterised by the closure of the AMP (in red, residues 28-72) and LID (in gold, residues 113-176) domains towards the CORE (in grey-blue) domain (Olsson and Wolf-Watz, 2010) around 8 hinges (circles labeled from 1 to 8, with hinge 8 located on the background helix (Henzler-Wildman *et al.*, 2007b))

4.2 Indentifying relevant community structures in AdK

The analysis proceeds as described in Chapter 3. We first convert the PDB file containing the crystal structure of *Escherichia coli* AdK (PDB 4AKE) into a weighted graph representation with edges based on identifying physico-chemical interactions. We then find partitions that optimise Markov stability at different Markov times.

The Markov stability analysis of AdK at all scales is shown in Figure 4.2. We first observe that as the Markov time increases, the optimal partition gets coarser: at very small values, each atom is identified as a distinct community; at very large times, the graph is partitioned into two large communities. This behaviour follows naturally from the definition of the Markov stability: with increasing Markov times, the probability that the diffusion process will stay within the smallest communities drops and larger communities become favored.

Secondly, at both small and large Markov times, it is apparent that certain partitions have long persistence, i.e. they remain optimal over long intervals of the Markov time. This persistence is a manifestation of their robustness: despite the process being allowed to explore the graph for longer, we cannot find a partition with larger communities where it is optimally trapped. These "plateaus" in the number of communities are thus an indication of the strong relevance of these partitions over the corresponding scales.

However, it is difficult to establish the significance of partitions in the intermediate regime of the Markov time. This is partly due to the fact that the number of possible partitions of intermediate size grows combinatorially. The larger ensemble of solutions renders the optimisation of Markov stability more difficult, and reduces the chances of always finding the single best partition.

In order to refine the evaluation of the robustness of the partitions, we calculate, at each Markov time, the average variation of information (VI) for an ensemble of 100 optimal solutions found starting from 100 random Louvain initial conditions and compare it with the VI of surrogate random graph models using a z-score statistic. The ensemble of surrogate models can be designed to test the null hypothesis. In this particular case, we use our knowledge of the intrinsic physico-chemical structure of proteins to formulate surrogates that can probe the emergence of biochemically relevant substructures at different scales. Indeed, the multiscale organisation observed in the case of proteins is particularly interesting because communities at different levels are linked to the presence of edges of different biophysical origins. For instance, the organisation of the protein in the form of a chain of amino acids is only defined by the network of covalent bonds, while higher levels of organisation, such as secondary, tertiary and quaternary structures, functional domains and rigid clusters, only depend on the position of the weak interactions and are essentially independent from the organisation of the covalent bonds outside of their role in maintaining the polypeptide chain. The biophysical origin of the different forms of structural organisations, which can either be chemical or spatial, leads to the definition of two types of surrogate random graph models for the robustness analysis.

4.3 Biochemically motivated null models

The normalised variation of information introduced in Chapter 3 does not give in itself an absolute value of the robustness of the partitions which is independent from the scale considered. Indeed, the number of possible partitions varies with the

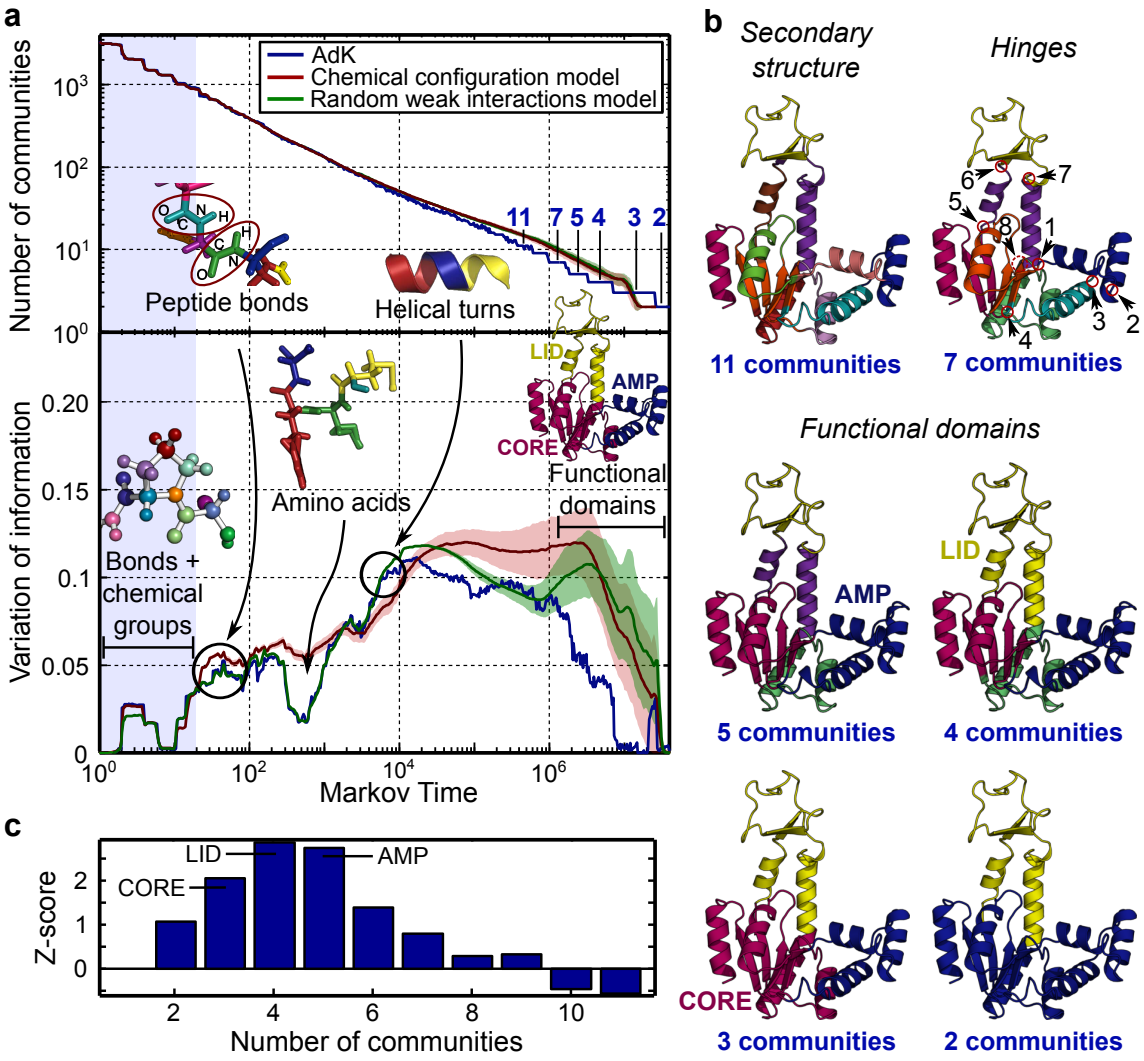


Figure 4.2: Markov stability analysis of AdK and comparison with biochemically motivated random graphs surrogates. **a.** Number of communities and variation of information as a function of the Markov time. Biochemically meaningful partitions at different scales (peptide bonds, amino acids, single helical turns, functional domains) show a very distinct robustness as compared to that of random graph models. The blue shaded area on the left corresponds to the Markov times for which the surrogates and the protein are equivalent, i.e. the level of the local chemistry which has been preserved in the random graphs. The green and red shaded areas around the curves correspond to one standard deviation. Real protein and randomized weak interaction surrogates diverge when weak interaction contribute to the formation of helical tunrs. **b.** Relevant partitions of AdK at large Markov times. The partition into 7 communities relates to previous hinge analyses (Meliga, 2009; Henzler-Wildman *et al.*, 2007a), and the 3-way partition to the functional domains. **c.** The z-score between AdK and the surrogates with randomised weak interactions indicate the partitions indentifying the functional domains as the most meaningful.

number of communities found, which changes with the Markov time. The values of the VI obtained for different scales can thus not be compared directly. This problem can be partly overcome by comparing the VI at each Markov time against a surrogate control group, obtained from a random graph model. The use of random graph surrogate models is a classical bootstrapping tool in graph theory (Newman, 2005). Here we use the z-score statistic to compare the robustness of the partitions of a particular graph with an ensemble of graphs from the random graph model,

$$Z(t) = \frac{VI(t) - \mu(t)}{\sigma(t)}, \tag{4.1}$$

where $\mu(t)$ and $\sigma(t)$ are the mean and standard deviation of the average VIs obtained for an ensemble of surrogate graphs generated from the random model. The Z-score can then be used as an estimate of the robustness of the partition which is independent from the number of communities detected.

4.3.1 Robustness at short scales: the chemical configuration model

Our first surrogate set is based on a random graph that preserves the local chemistry of the protein while randomising all other interactions. This can be used as a chemical null model that should be identical to our original graph at short time and length scales but will highlight the differences that emerge with the larger scale organisation. The random graph model is designed to preserve the basic chemical attributes of the protein including its chemical composition by preserving the valence of the atoms, encoded in the degree of the nodes, and the energies of the bonds and interactions, encoded in the weights of the edges. All the basic chemical properties of the graph can be kept using a simple randomisation scheme similar to the one proposed by Maslov and Sneppen (Maslov and Sneppen, 2002), in which pairs of bonds chosen at random exchange one of the two nodes they link. By doing this repeatedly, a new random graph keeping the number but also the weights of the connections of each node is generated. The same method is used here, with two additional constraints: Firstly, the pairs of bonds to be exchanged must be of the same kind (covalent bonds of the same energy, or weak interactions of the same nature), and secondly, the exchange must keep the whole network of covalent bonds connected within each monomer (the number of subunits is maintained). This randomisation thus also keeps the chemical nature of the neighbours of each atom. Consequently, from a chemical point of view, the small chemical groups are kept,

and, from a graph theoretical point of view, the degree of each node is also maintained. In that respect, this model is similar to the configuration model (Molloy and Reed, 1995) and can be thought of as the "chemical configuration model".

4.3.2 Robustness at long scales: randomised weak interactions

The large scale spatial organisation of the protein is mainly determined by the weak interactions such as hydrogen bonds, hydrophobic tethers, salt bridges, electrostatic and pi-stacking interactions. The second type of surrogate random graph conserves the whole network of covalent bonds defining the primary structure of the protein, but randomises the positions of the weak interactions which govern the secondary and tertiary structures. The randomisation of these interactions is carried out preserving the necessary chemical constraints: hydrogen bonds should only bind oxygen or nitrogen with hydrogen atoms and hydrophobic tethers, carbon and sulphur atoms. The weak interactions are then re-positioned between nodes of the required nature selected at random.

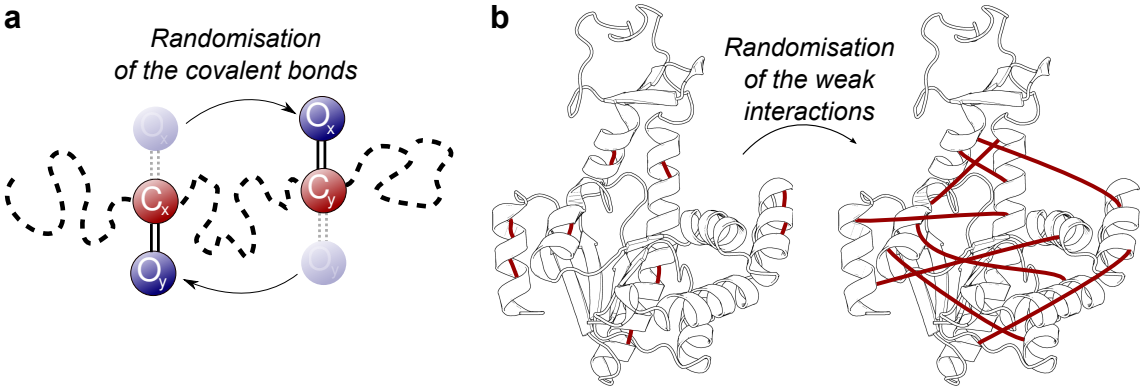


Figure 4.3: Illustration of the randomisation schemes used. **a.** Covalent bonds are randomised in the chemical configuration model by repeatedly swapping nodes from covalent bonds of the same types. **b.** Weak interactions are randomised by rewiring at random each subnetwork formed by the weak interactions of each particular type.

4.4 The Markov stability analysis of AdK

Results of the partitioning and robustness analysis for AdK are summarized in Figure 4.2. At each Markov time, the Markov stability was optimised a hundred times with different Louvain initial conditions. For each Markov time, the number of communities of the optimal partition is shown in the top panel of Figure 4.2a, and

the variation of information between all the partitions found at this Markov time is shown in the bottom panel. Partitions at very small and very high Markov times remain optimal for extended periods of Markov time and correspond to biochemically meaningful components such as small chemical groups (small times) or the three functional domains (LID, AMP and CORE domains). This is confirmed by our robustness analysis, which shows small values of VI for the long-lived partitions.

Figure 4.2 also shows the comparison of the robustness of the partitions of the protein against that of ensembles of random graphs from our surrogate models. As expected, the random graphs obtained from the chemical configuration model are indistinguishable from the protein at short Markov times, since their local chemical structure is identical. However, at longer times the comparison reveals two additional partitions of strong biochemical significance corresponding to the peptide bonds between the amino acids (at Markov times around 30) and to the emergence of amino acids at Markov times around 500. At the local minimum of VI, 63% of the amino acids were grouped as a community, while most of the others, essentially small amino acids, were grouped with another residue.

The robustness of the protein is indistinguishable from the ensemble of graphs obtained by randomization of weak interactions until Markov times of around 6000. This establishes the spatial and time scales at which the weak interactions start having an influence on the communities, and, by extension, on the conformation of the protein in space. Interestingly, the communities identified at this Markov time correspond to the helical turns, which can indeed be thought of as the smallest biochemical building block at which weak interactions start to play a role in the structural organisation of the protein. This is another demonstration that the parametrisation established in the previous chapter yields the expected behaviour in the Markov stability analysis.

As expected, the two random models converge at long Markov times since in both cases the composition of the molecule is conserved, and the weak interactions have been randomized. At Markov times above 10^5 we observe an increase in the variability and a decay in the value of the VI for the surrogates from the weak interaction randomization. This indicates the point where weak interactions placed at random begin to induce robust compact subgroups in the structure in an effect akin to undirected packing. In contrast, the specific location of the weak interactions in the structure of the protein induces robust and reproducible partitions that reveal the specific organisation of the protein conformation.

At long Markov times, Markov stability finds partitions into a few subunits that are much more robust for the protein than for the random surrogates. The study of their robustness indicates the relevance of partitions into two, three, four and five communities. Interestingly, each of these partitions corresponds to the identification of one of the three functional domains. The AMP domain is fully clustered into one community for the first time in the five-community partition, followed by the LID domain in the four-community and finally the CORE domain in the three-community partition, at which point the three functional domains correspond to the three communities and the variation of information drops at zero.

On Figure 4.2, three sharp drops in the variation of information can also be noticed at Markov times 2×10^6, 3×10^6 and 7×10^6. Each drop corresponds to the identification of a new partition (marked by a new plateau in the number of communities), into five, four and three communities respectively, which each mark the detection of one functional domain by Markov stability. These sudden decreases in the variation of information thus reveal the "crystallisation" of one region of the graph into a robust community, which thus ceases to contribute to the total variation of information observed at the previous Markov time point. We will see in the next chapter that this observation holds in other structures as well, which will allow us to circumvent the need to compute the surrogate random graphs and z-score statistics.

Although the partitions into seven and eleven communities do not distinguish themselves from the surrogates in terms of their robustness (see z-scores on Figure 4.2), both have a plateau in the number of communities, and correspond to well known levels of organisations in the structure of AdK. The partition into eleven communities almost perfectly identifies the secondary structure, with each individual helix, and each individual β-sheet englobed into a distinct community. The partition into seven community is highly correlated with the location of the eight hinges around which the LID and AMP domains rotate upon closure of AdK (Müller *et al.*, 1996; Henzler-Wildman *et al.*, 2007b). Of the eight hinges, only hinge 2 is not located at the frontier of two communities.

4.5 Closed form of AdK

The discovery that AdK does actually explore its closed conformation spontaneously, even in the absence of substrate has been at the origin of a new understanding of the functionning of enzymes, and led to the hypothesis that the spontaneous closure could be a more widespread property.

In Figure 4.4, we compare the Markov stability of four structures of *Escherichia Coli* AdK in both open and closed conformations. The similarity of our results for all four structures in spite of the huge difference in their conformation (RMSD of 7.4 Å between the two conformations, see Figure 4.1) is striking, and shows the ability of our analysis to capture this fundamental property of AdK. Indeed, the absence of any difference in the structural organisation of AdK as identified by the Markov stability analysis suggests that the closed conformation is not the result of a structural change induced by the ligand, but an intrinsic property of AdK in any conformation.

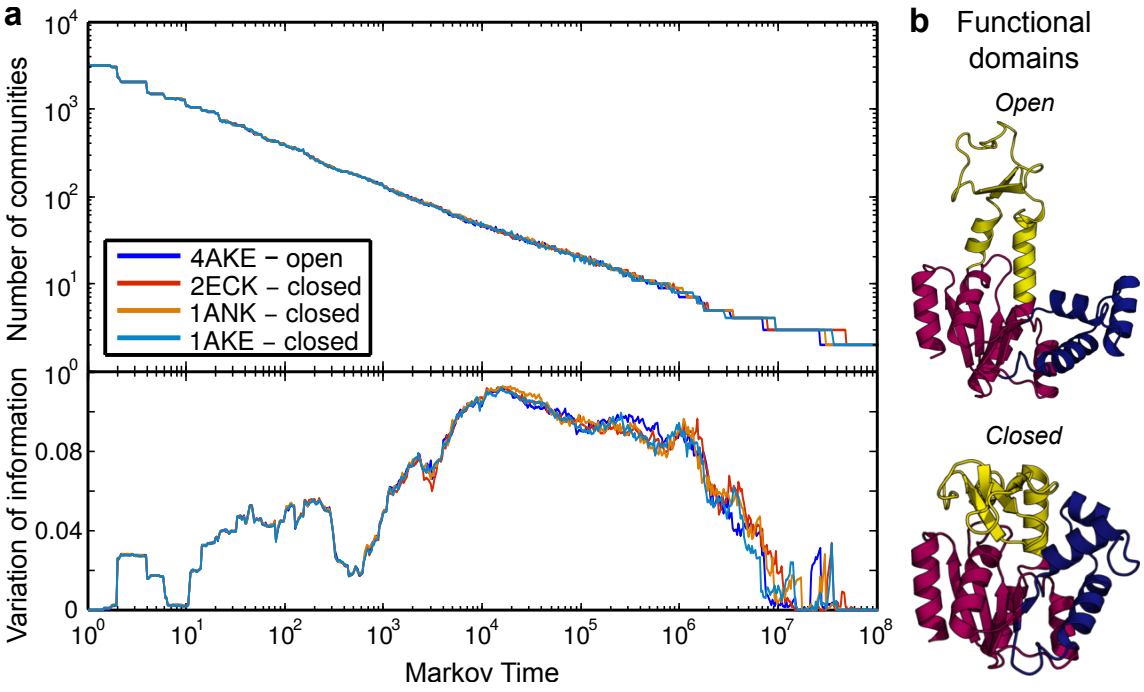


Figure 4.4: Markov stability analysis of open and closed conformations of AdK. **a.** The Markov time evolutions of the number of communities and variation of information of AdK in the open and closed conformations are almost indistinguishable. Both forms are known to coexist in solution, and AdK and other enzymes are thought to close spontaneously in the absence of substrate (Arora and Brooks III, 2007; Henzler-Wildman *et al.*, 2007b). The structural organisation captured by Markov stability is thus an intrinsic property of the protein rather than of a particular conformation. **b.** The biologically meaningful communities – here the functional domains – are almost identical in both structures.

4.6 Discussion

In this chapter, we demonstrated the suitability of the methodology introduced in Chapter 3 for the analysis of biomolecular structures through its application on a classical and well documented protein example, *Escherichia coli* adenylate kinase. In particular, we showed that the chemical, biochemical and biological levels of organisation detailed in the first chapter, i.e. chemical groups, residues, secondary structures and functional domains (see Figure 1.1), are all successfully recovered by the method at increasing scales, and in so doing, confirmed results previously obtained by Meliga (2009). Communities identified through the partitioning algorithm at a particular Markov time relate to regions of the protein sharing common dynamical properties over a certain range of time scales. Two sets of biochemically-motivated random graph models were also introduced in this chapter against which we tested the significance of our results. Beyond providing a benchmark against which our robustness analysis could be tested, they also demonstrated the validity of our methodology by identifying in the real protein biochemically meaningful structural levels of organisation that significantly diverged from the random graph surrogates such as peptide bonds, residues, helical turns and functional domains. Finally, our analysis of the open and closed forms of AdK revealed a surprisingly high similarity between the two structures in spite of their largely different conformations. While hinges and functional domains are expected to be conserved accross conformations, our observation goes beyond: It is the entire structural organisation that encodes the dynamics of the whole protein which is strictly conserved. The experimental observation that AdK explores both conformations spontaneously in the absence of substrate fully aligns with our results.

The possibility of a relation between the Markov time at which communities are detected and the time scale of motion of their corresponding region in the protein has been previously hypothesised by Meliga (2009). From Figure 4.2, it indeed appears that the communities identified at increasing Markov times are linked to increasing time scale of motions: Individual covalent bonds are first found, followed by chemical groups, then residues, secondary structure elements and finally the functional domains, which corresponds to their ordering in Figure 1.1.

We find this observation to carry over to the time scale of motion of the individual domains in AdK. In our analysis, the AMP domain is the first to cluster, followed by the LID and finally the CORE domain when the Markov time is increased. This is consistent with results from a Molecular Dynamics and Normal Mode Analysis

study of AdK (Lou and Cukier, 2006) which reports the three slowest normal modes of AdK to be linked to domain motions. The fastest mode 3 is dominated by the movement of the AMP domain, and the slowest mode 1 by the movement of the LID domain while mode 2 describes a collective motion of the LID and AMP domains. The three slowest normal modes do not show a collective motion in the CORE domain, and it also only forms a community in our analysis after the AMP and LID domains have been found. Although the generality of this observation remains highly speculative, the time scale of motion of each of the three domains of AdK appears to be in agreement with the Markov time at which each of them is identified by our analysis.

Chapter 5

The myosin tail interacting protein and mutational analysis

ADENYLATE kinase provided an ideal case study to test our framework and exemplify its capabilities and its limitations. We now use our methodology to contribute towards understanding the functioning of the myosin tail interacting protein (MTIP), a recently discovered anchoring protein from a myosin A molecular motor complex whose structural organisation and dynamics are still poorly understood (Bergman, 2002).

In this chapter, we investigate the structure of MTIP through the lens of Markov stability and identify regions of the protein sharing a common dynamical behaviour at a particular scale, leading to a hypothetical closing mechanism. The same tool is then used to explain the differences observed between different conformations and between structures from two different species.

Finally, we introduce a measure based on Markov stability which estimates the impact of a particular residue on the global structural organisation of the protein and thus suggests key residues or "hotspots" that could be targeted through mutagenesis. We subsequently use it to probe each amino acid of the tail of myosin A, the motor protein binding MTIP.

5.1 MTIP and myosin-myosin light chain interactions

Most forms of movement in living organisms are implemented by motor proteins (Schliwa and Woehlke, 2003). The most prominent example is myosin II which, powered by the hydrolysis of ATP, carries out the contraction of muscle cells by

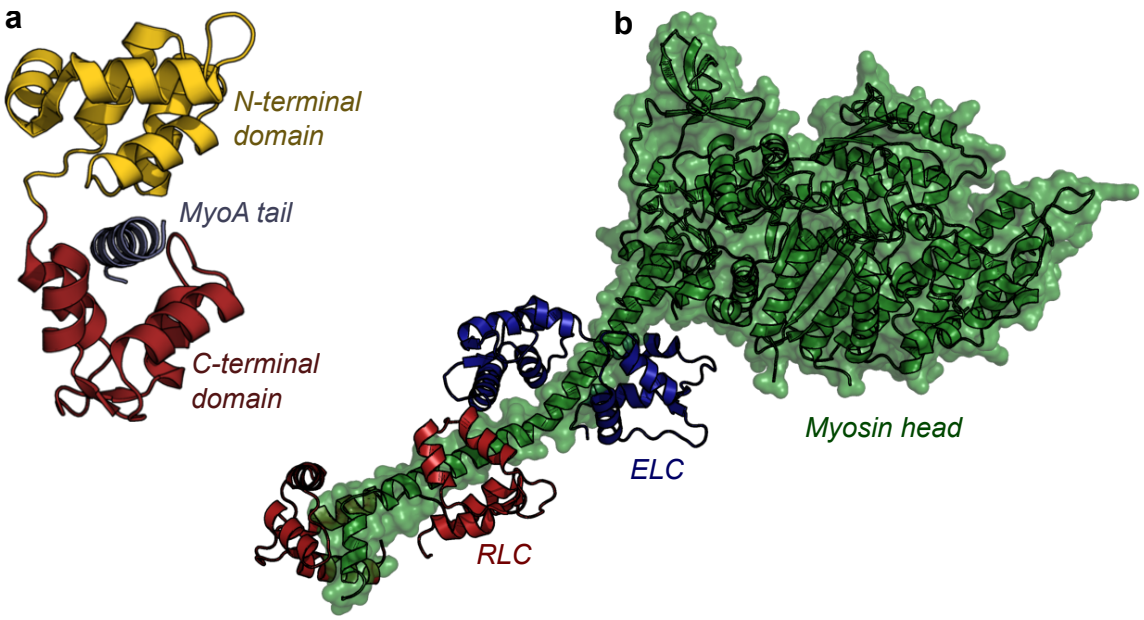


Figure 5.1: **a.** Structure of *P. falciparium* MTIP in complex with the MyoA tail (PDB 4AOM). **b.** Crystal structure of the scallop muscle myosin essential (ELC) and regulatory (RLC) light chains in complex with the myosin heavy chain (PDB 1QVI). This ELC is the closest structural MTIP homolog (Bosch *et al.*, 2006) and, when bound to the myosin heavy chain, adopts a conformation similar to the way MTIP wraps around the MyoA tail.

pulling against actin filaments (Eisenberg and Hill, 1985). All myosin forms consist of a long tail terminated by a head region. The head is usually subdivided into the actin-binding motor domain activated by ATP and the neck region formed by a single α-helix . The latter acts as a lever arm and also serves as the binding domain for calmodulin-like molecules called myosin light chains (Figure 5.1b) (Sweeney and Houdusse, 2010; Lowey and Trybus, 2010; Schiaffino and Reggiani, 1996).

Myosin light chains, such as the essential and regulatory light chains in mammalian muscle cells, play a major role in regulating and fine-tuning molecular motor complexes, notably by regulating ATPase activity and actin binding affinity, and even by interacting directly with the actin filaments (Trybus, 1994; Timson, 2003). Crystal structures of the myosin head and neck regions (Rayment *et al.*, 1993) also suggested a possible role of the light chains in stabilising the lever arm, thereby allowing for a more powerful stroke in the cross-bridge cycle[1].

Here we focus on the myosin tail interacting protein (MTIP) (Bergman, 2002), a myosin light chain analog forming part of the molecular machinery which allows

[1]The sequence of reactions generating the motion.

Plasmodium species to invade red blood cells. *Plasmodium* species are the causative agents of malaria, an endemic disease affecting most developing countries with an estimated 500 million people being contaminated and three million killed every year (Snow *et al.*, 2005). The life cycle of the malaria parasite includes a series of development phases in the *Anophele* mosquito and human. The human blood phase involves the invagination of red blood cells by the parasite using an acto-myosin motor system (Baum *et al.*, 2006; Farrow *et al.*, 2011; Besteiro *et al.*, 2011) based on an unconventional class XIV myosin, called myosin A (MyoA) (Heintzelman and Schwartzman, 1997; Hettmann *et al.*, 2000), which lacks the C-terminal tail. MyoA is anchored inside the parasite via the MTIP protein to an inner membrane complex located just behind the plasma membrane (Bergman, 2002; Rees-Channer *et al.*, 2006; Green *et al.*, 2006; Frénal *et al.*, 2010). This protein-protein interaction is key to this stage in the life cycle of the parasite, and therefore to its survival, and is consequently increasingly being seen as a potential target for the design of new anti-malarial drugs which could overcome the typical drug resistance effects in malaria (Kortagere *et al.*, 2010; Douse *et al.*, 2012).

The structural organisation of MTIP and the ways in which it interacts with the MyoA tail are still poorly understood. A first crystal structure containing three conformations of *Plasmodium knowlesi* MTIP (PkMTIP) suggested that MTIP binds MyoA via the two lobes of the C-terminal domain wrapping around the MyoA tail (Bosch *et al.*, 2006) (Figure 5.1a). Subsequent crystal structures of *Plasmodium falciparium* MTIP (PfMTIP) and binding assays showed the N-terminal domain to influence binding as well, and suggested the existence of a conformational change which would allow the N-terminal domain to bind with the MyoA tail directly (Thomas *et al.*, 2010; Bosch *et al.*, 2007). MTIP was initially suggested to only bind the last 15 residues of the MyoA tail (Bergman, 2002), but recent experiments (Thomas *et al.*, 2010) showed the interaction to be much stronger when the last nineteen residues are included in binding assays.

Further developments in anti-malarial drugs targeting MTIP or the motor complex require first a deeper understanding of the functional organisation of MTIP and particularly its binding mechanism to the MyoA tail. The goal of this analysis is to investigate further the mechanism by which MTIP wraps around the MyoA tail by understanding the changes that occur in the structural organisation of MTIP upon binding, and to identify amino acids of the MyoA tail that play a key role in this mechanism. To this end, we use Markov stability to study the multiscale structure of MTIP and evaluate how different parts of the protein behave together over dif-

ferent scales. By comparing our results for different conformations, we explore how this functional organisation changes upon binding. Finally, we introduce a methodology to measure the impact of individual residues on the multiscale structure which allows us to identify key binding residues of the MyoA tail.

5.2 Structural data

We here study in detail four crystal structures of *Plasmodium knowlesi* (Pk) and *Plasmodium falciparium* (Pf) MTIPs, either unliganded or in complex with a MyoA tail peptide. The PkMTIP structures (PDB 2AUC) have been obtained by Bosch *et al.* (2006) through X-ray crystallography at 2.6 Å resolution and pH 5.3. The asymmetric unit in the crystal shows MTIP residues S803 to A817 and comprises three conformations: two unliganded forms and one structure in complex with residues S803 to A817 of *P. yoelii* MyoA tail. The crystal structure of PfMTIP (PDB 2QAC) has been resolved by the same group (Bosch *et al.*, 2007) at 1.7 Å resolution and pH 7.5, and comprises MTIP residues S61 to Q204 in complex with the same 15-residue *P. yoelii* MyoA tail peptide.

For each structure, we performed two preliminary energy minimisation steps using the molecular dynamics package Gromacs (Hess *et al.*, 2008) with the steepest descent and conjugate gradient algorithms. After having immersed each protein in a cubic box with spc216 water molecules and periodic boundary conditions, ensuring that the borders of the box are at least 10 Å away from the protein at all points, the two energy minimisations were done using the GROMOS96 43a1 force field (Christen *et al.*, 2005) until convergence.

5.3 Connections to a rigid cluster and the closing mechanism

We now use our Markov stability framework to study the structural organisation of PfMTIP/MyoA[803-817], i.e., PfMTIP in complex with a peptide of the last 15 amino acids of the MyoA tail (PDB 2QAC).

At small Markov times (Figure 5.2), we find partitions of high robustness (large z-score) corresponding to the peptide bonds and individual residues, similarly to our analysis of AdK.

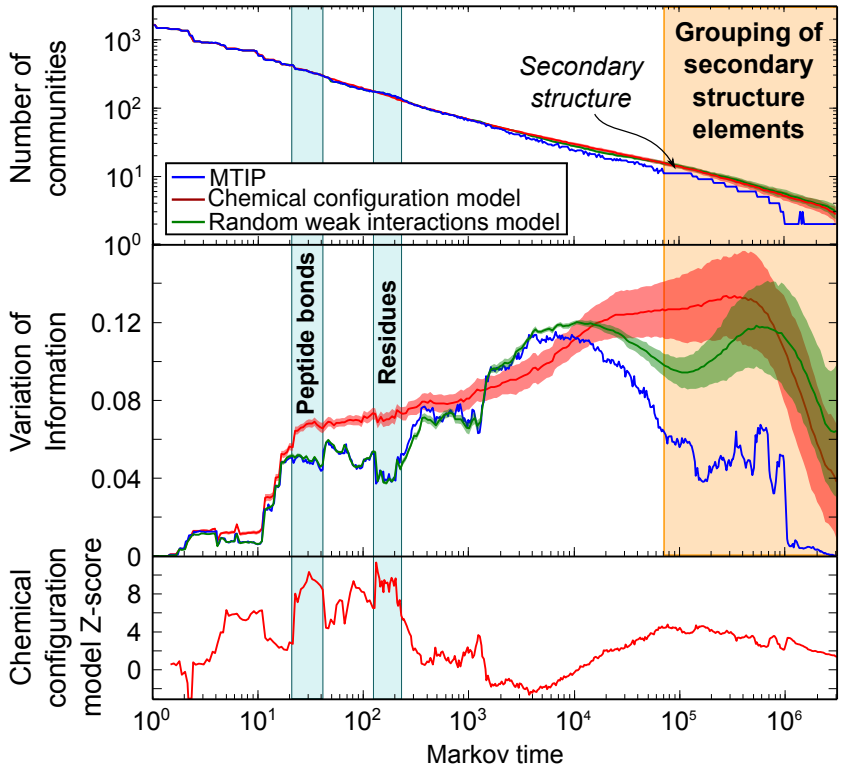


Figure 5.2: At small Markov times, residues and peptide bonds (blue shaded regions) are correctly identified as relevant communities by the z-score statistics (bottom panel) contrasting the robustness of the PfMTIP partitions to that of random graphs from the chemical configuration model across Markov times. At large Markov times, the results for PfMTIP substantially deviate from both the chemical configuration and random weak interactions models indicating the presence of relevant partitions. The shaded area in orange, corresponding to the scales beyond the secondary structure, is further analysed in Figure 5.3.

The relevant partitions at long Markov times are summarized in Figure 5.3a. Starting from the detection of the secondary structure at Markov time 8×10^4, the different α-helices and β-sheets are progressively incorporated in a quasi-hierarchical manner into bigger clusters as the Markov time increases. Some of the groupings lead to a marked increase in the robustness of the partition (indicated by the higher z-score in Figure 5.3a & c). This is the case for the first community to appear that incorporates two secondary structure elements: helices α6 and α7. This community is conserved across a broad range of Markov times, more than any other community of multiple elements of secondary structure. Following from our results in AdK, this suggests a strong dynamical linkage between these two α-helices over an extended time scale of motion. This result is in agreement with previous analysis of the PkMTIP crystal structure by Bosch *et al.* (2006), which suggested that these helices

form together a rigid cluster. Indeed, in all three conformations present in their crystal, comprising both liganded (MyoA-bound) and unliganded (free) structures, these two helices always keep their relative position unchanged.

Another important community is formed by helices $\alpha 5$ and $\alpha 8$ and similarly leads to a marked increase in the robustness of the partition. Together with the $\alpha 6 - \alpha 7$ cluster, they divide the C-terminal domain into two regions corresponding to the two lobes that wrap around the MyoA peptide. The next rise in the z-score appears at Markov times 7×10^5, at which point MTIP is divided into three domains: the two lobes of the C-terminal domain and the entire N-terminal domain (Figure 5.3b). The strong robustness of this particular partition reflects its significance for the functioning of the protein. This again supports hypotheses from Bosch *et al.* (2007) that the closing mechanism of MTIP around MyoA should be in the form of a clamp, with the two lobes of the C-terminal domain wrapping around the MyoA tail, and the N-terminal domain fortifying the binding by bending towards the C-terminal domain to close the clamp.

Strikingly, helix $\alpha 0$ forms another very long-lived community and remains dissociated from the rest of the MTIP structure for almost as long as the MyoA peptide—which does not form any covalent bond with MTIP. Preceding the sequence MTIP[61-204] resolved in the structure we analysed, 60 additional residues form part of a third unresolved domain (Frénal *et al.*, 2010) through which MTIP binds the protein GAP45 and thus anchors itself to the inner membrane complex of the malaria parasite. The strong separation of helix $\alpha 0$ from the rest of the N-terminal domain as found by Markov stability suggests that this helix could in fact be the only resolved portion of this third domain in this structure.

At long Markov times, the complex is partitioned into N- and C-terminal domains, with the MyoA peptide clustered with the C-terminal domain. This is also in agreement with results from K_d analyses (Thomas *et al.*, 2010), which suggest that the MyoA tail should be more tightly bound to the C-terminal than to the N-terminal domain.

5.4 Stabilising role of MyoA and similarities between conformations

We here study the changes in the structural organisation of MTIP induced by the presence of the MyoA peptide by comparing our results for unliganded and liganded

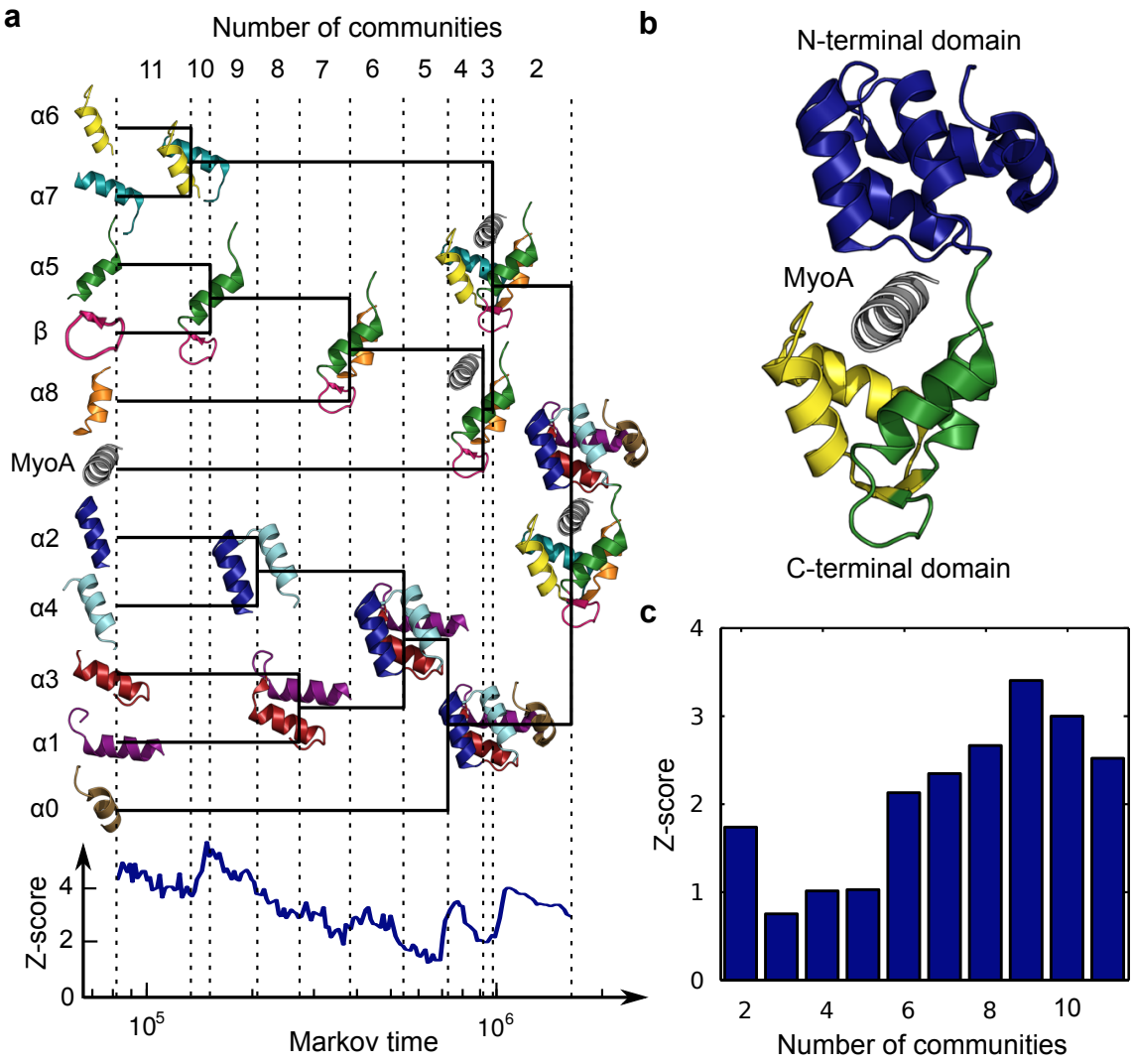


Figure 5.3: **a.** Multi-scale partitioning of PfMTIP/MyoA as a function of Markov time. The elements of the secondary structure are progressively grouped into larger communities as the Markov time evolves. Although in general our methodology does not pre-impose a hierarchical community structure, in this case the succession of community groupings is close to a strict hierarchy. Clusters kept for a long range of Markov times, such as the group of helices $\alpha 6$ and $\alpha 7$ are well-defined partitions. The identification of the rigid cluster (ten communities) and functional domains (four communities) leads to an increase in the robustness (z-score) of the partitions and a drop in the variation of information. **b.** Detection of the functional domains in the four-community partition of PfMTIP/MyoA. **c.** The comparison of the z-score per number of communities suggests that the partitions into nine and ten communities, where the rigid cluster is found, and the partition into two communities, where the N and C-terminal domains are identified, are significant.

structures of MTIP. When generating the graph prior to the analysis, the MyoA peptide and all its interactions with MTIP have here been removed from the liganded structures to make the results more comparable. In doing so, the graphs constructed from liganded structures reflect solely the change of conformation in MTIP induced by the MyoA peptide, independently from the effect of the direct constraints, and the analysis is thus focused on MTIP only.

Figures 5.4a & b show that the partitions for the liganded structures (PkMTIP3 and PfMTIP) obtained from the complexed forms are in general much more robust than the partitions for the two unliganded structures (PkMTIP1 and PkMTIP2), especially at the level of the secondary structure (eight, nine, and ten communities) and functional domains (three communities). Such increase in the robustness of the partitions in the liganded conformations thus emerges naturally from the change in the spatial structure induced by the MyoA peptide.

Importantly, although the partitions differ significantly in their robustness and the Markov time of their predominance, they are themselves very similar among the different conformations, especially between the two liganded forms. In particular, the important communities identified in the previous section, such as the $\alpha 6 - \alpha 7$ cluster and the functional domains (Figure 5.3b), are also detected in all three conformations with high robustness (except for the functional domains of PkMTIP2 whose robustness is comparatively low). The fact that the same partitions are found in all structures is in line with our previous results on AdK in closed and open forms. This suggests that the overall organisation of MTIP, and particularly the aspects of it tightly linked to function, is maintainted throughout the different conformations. This observation is in line with the previously proposed hypothesis that the structure of proteins could have been optimised by evolutionary selection for an efficient exploration of their conformational space (Henzler-Wildman *et al.*, 2007b,a; Henzler-Wildman and Kern, 2007): the functionally important conformations are already encoded in the fold.

The changes in the robustness and Markov lifetime of the partitions however suggest that the secondary and tertiary structures get better structured upon binding with the MyoA tail since the corresponding partitions are better defined in this case. Note also that in comparing the liganded conformations, PfMTIP has more robust partitions than PkMTIP3, possibly a result of the stabilising role of the N-terminal domain, which in PfMTIP also binds the peptide and closes the clamp. On the other hand, the unliganded form PkMTIP2 possesses the least robust partitions, which is

in accordance with the hypothesis proposed by Bosch *et al.* (2006) that it should be an intermediate conformation between the fully opened and fully closed forms.

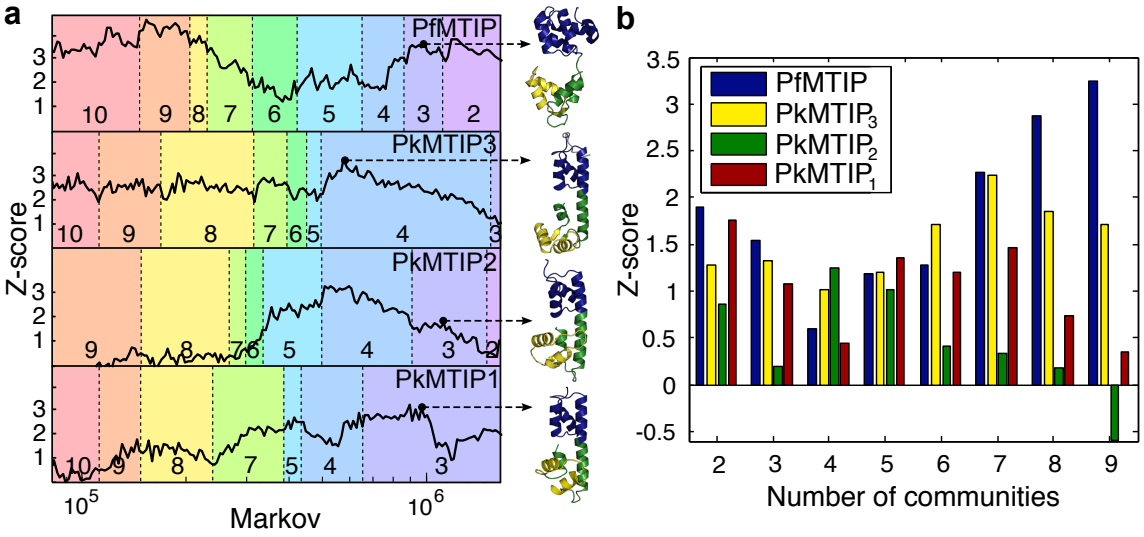


Figure 5.4: **a.** Robustness of the partitions of MTIP in different conformations as a function of the Markov time. The liganded conformations (PfMTIP and PkMTIP3) show better properties of robustness than the unliganded ones at the level of the secondary structures and of the functional domains, suggesting a stabilising role of the binding with MyoA. Partitions are very similar between the three conformations, in particular for the functional domains, although the grouping of helices α5 and α8 only occurs at long Markov times for PkMTIP2. Note: PkMTIP3 has four communities instead of three for the functional domains due to helix α0 being only partially resolved and thus failing to fully merge with the rest of the N-terminal domain. **b.** The z-score of the partitions with the same number of communities compared across different conformations of MTIP shows that the liganded forms PfMTIP and PkMTIP3 have better defined partitions in general.

5.5 Robustness of the secondary structure

In spite of a very high sequence similarity (80%), the conformations adopted by PfMTIP and PkMTIP differ considerably, with a RMSD of 1.06 Å for 121 amino acids. While the N and C terminal domain are linked by an extended α-helix in PkMTIP, PfMTIP adopts a kinked conformation which allows the N-terminal domain to interact with MyoA (Figure 5.4a). Considering the unphysiological pH of 5.3 at which the PkMTIP structures were obtained, their structural differences with other myosin light chains, as well as the smaller construct (portion of the MTIP

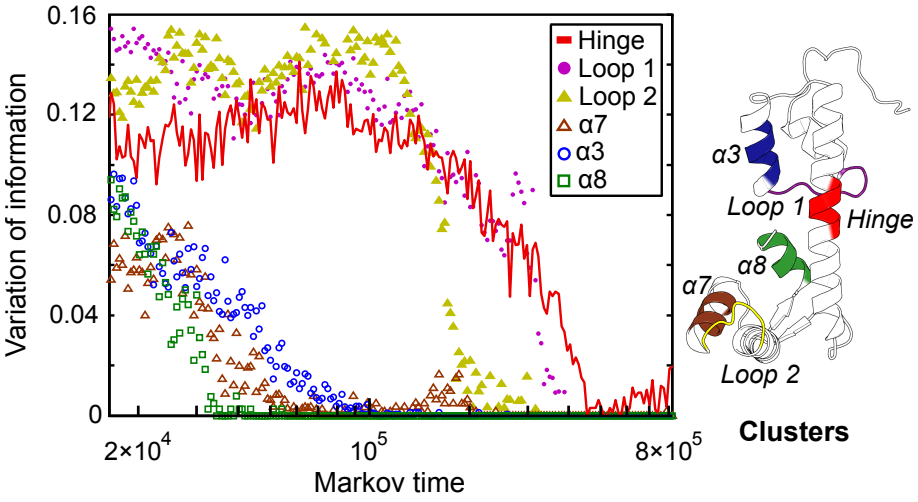


Figure 5.5: Variation of information of the partitions detected in different secondary structures. The variation of information of the central region of the central α-helix (continuous red line) is very high and behaves similarly to other loop regions of the protein. This suggests that the algorithm effectively recognizes this region as a loop, despite its α-helical secondary structure. Filled symbols correspond to loops, empty symbols correspond to α-helices and the continuous line corresponds to the hinge region.

sequence) used to obtain the crystal structure, Bosch *et al.* (2007) suggested that PkMTIP also adopts a compact conformation similar to the PfMTIP structure in physiological conditions.

The similarity between the partitions of the complexed forms of PfMTIP and PkMTIP (PkMTIP3) obtained in the previous section also supports the expectation that their structural organisation should not be very different. Our partitioning indeed consistently divides the central α-helix of PkMTIP into two different communities at all Markov times. Furthermore, the separation between the two halves of this central α-helix in the partitions is constrained within the region that corresponds to the central loop in PfMTIP (from residues H135 to N140). This partitioning is thus consistent with the central α-helix of PkMTIP being partly identified as a loop by the partitioning algorithm. To further support this observation, we have carried out an analysis of the robustness of loops and α-helices with the same number of nodes (50 atoms) across Markov times. Figure 5.5 shows that the central region of the central α-helix of PkMTIP has a robustness much lower than the typical α-helix with a profile similar to that of loops. These results demonstrate the insights that our method can bring into the analysis of the structural organisation of a protein beyond its pre-assigned secondary or tertiary structure.

5.6 MyoA tail computational mutational analysis

The last part of the analysis aims at identifying residues in the MyoA tail that have a strong impact on the multiscale organisation of the protein complex and can therefore be considered to play a significant role in its structure and dynamics. This analysis does *not* evaluate the influence of a mutation on the binding energy; rather, the expectation is that residues with a large influence on the structural organisation of the protein will affect the global dynamics of the binding events.

Typically only a very small number of residues contributes to most of the binding (Clackson and Wells, 1995). These residues, coined "hot spots", are commonly defined as those whose mutation to alanine produces a change in binding free energy of 2.0 kcal/mol or more (Bogan and Thorn, 1998). Alanine scanning mutagenesis is the standard experimental method to identify hotspots. Each residue to be analysed is sequentially converted to alanine and the change in the binding affinity is then measured (Wells, 1991). Although mutations to different amino acids can be considered, alanine has the advantage of minimising the side chain without adding extra flexibility to the backbone. Alanine is also the most common amino acid in proteins, and is found in all secondary structures, in buried as well as non buried regions.

Although most computational methods to find hotspots are naturally energy-based (Moreira *et al.*, 2007; Morrow and Zhang, 2012), functionnally critical residues are often linked to the global mechanical properties of the protein, and experimental evidence has associated them with flexibility and intrinsically disordered regions (Ma *et al.*, 2001; Radivojac *et al.*, 2007; Henzler-Wildman *et al.*, 2007a; Costa and Yaliraki, 2006). Furthermore, various computational methods have demonstrated the high influence of binding site residues on large-scale attributes such as the distribution of conformations (Ming and Wall, 2006), the network of cooperativity between residues measured in terms of coupled fluctuations (Liu *et al.*, 2007), or their propensity to be located in regions with distinctive mobility patterns in the slow modes (Yang and Bahar, 2005).

To assess whether the contribution of each residue to the binding energy is related to its impact on the structural and dynamical features detected by our method, we have compared our computational results with the outcome of binding assays of MTIP with mutated MyoA tail peptides. Our computational setup mimics the standard alanine scanning mutagenesis experimental procedure: Each residue is "mutated" in turn by removing from the graph all the edges corresponding to the

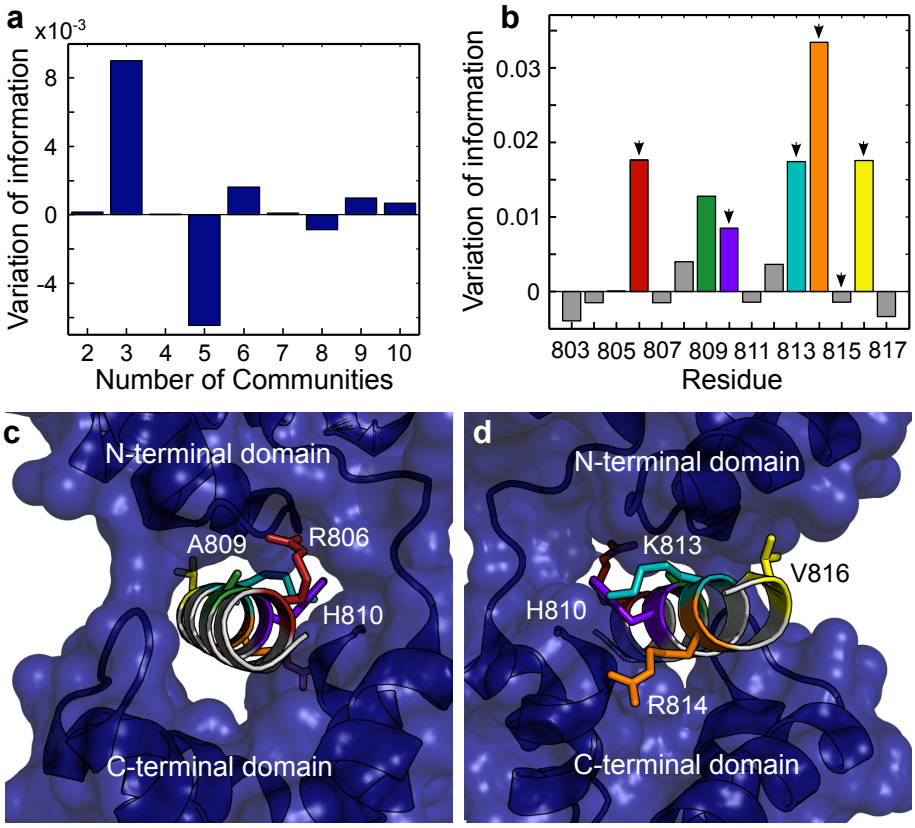


Figure 5.6: **a.** The partition most influenced by computational mutagenesis is the one into three communities. **b.** Residues R806, A809, H810, K813, R814, and V816 have the biggest influence on the three-way partition. There is a strong coincidence with the residues identified experimentally (Thomas *et al.*, 2010) to have a strong effect on the binding affinity (indicated by the black arrows). **c & d.** Front and side view of the positions of the key residues found.

weak interactions it makes with other residues . The mutated graph is then analysed with our multiscale methodology, and the partitions are compared with those of the original graph using the VI. For each mutation, we compute the VI between all the partitions found with the same number of communities from the original and mutated protein averaged over ten different Louvain initial conditions from which we substract the average VI of the original graph to renormalise the results. Using this scheme, partitions which are the most affected by a particular mutation will give a high value of the variation of information.

Figure 5.6a shows that the partitions into three communities are the most affected by the mutations. This is not surprising since the three-way partition is the first where the MyoA peptide is grouped with part of the MTIP molecule (Figure 5.3a). Consequently, the mutations essentially affect the strength of the asso-

ciation between the MyoA tail and the portion of MTIP that includes the hinge region and the helices $\alpha 5$ and $\alpha 8$ from the C-terminal domain. More specifically, the mutations that cause the largest changes in the three-way partition are those in residues R806, A809, H810, K813, R814, and V816 (Figure 5.6b). These results are in accordance with experimental binding assays for MyoA peptides of different lengths (Thomas *et al.*, 2010), crystallographic data and yeast two-hybrid experiments (Bosch *et al.*, 2007). In particular, residues R806 and K813 have been observed to be essential for complex formation; H810 and R814 provide key contacts for tight binding; and V816 also generally improves the binding strength. On the other hand, our method does not single out a significant contribution of residue M815, which has also been found experimentally to influence binding affinity. A possible explanation is that the importance of this residue might be related to effects that are not directly addressed by our method, such as intermediate states in the folding pathway or a pure modification of the binding energies without any other impact on the structure and dynamics of the complex. Finally, our method finds one residue, A809, predicted to have an important effect on the multiscale organisation which has not been investigated experimentally to date. This residue is however one of the only four completely conserved residues of the MyoA tail in *Plasmodium* species and other affiliated apicomplexan parasites (Bosch *et al.*, 2007).

5.7 Conclusion

Our analysis has uncovered important features of the MTIP/MyoA complex that agree well with experimental data. The rigid cluster formed by helices $\alpha 6$ and $\alpha 7$, as suggested by the crystal structures of PkMTIP (Bosch *et al.*, 2006), was observed to form a well-defined community, conserved across a broad range of Markov times and associated with very robust partitions. The functional domains suggested by the analysis of crystal structures of different conformations across species (Bosch *et al.*, 2006, 2007) have been detected by the partitioning and also showed strong robustness and conservation across Markov times. The robustness analysis of the hinge region of PkMTIP confirms these similarities between species and therefore suggests that their dynamical behavior should be similar. Furthermore, it supports the hypothesis (Bosch *et al.*, 2007) that the reported differences between PkMTIP and PfMTIP in the the hinge region could result from the particularities of the crystallisation. Finally, a computational tool for mutational analysis was introduced and used to identify five out of the six residues known from binding assays (Thomas

et al., 2010) to have a strong influence on the binding of MyoA. It also suggested one additional residue, A809, which has not yet been investigated experimentally, to be particularly important.

The experimental identification of hotspots through alanine scanning mutagenesis is a slow and laborious task. Consequently, a variety of computational methods have been developed in the past, notably using free energy calculations, molecular dynamics simulations, or machine learning (Fernández-Recio, 2011; Moreira *et al.*, 2007; DeLano, 2002; Morrow and Zhang, 2012). The mutational analysis tool we introduced in this chapter is not aimed at competing in accuracy with these methods, but rather offers a different perspective to the identification of key residues for the functioning of a protein. Our graph theoretical approach allows for instance to account for indirect or long-range effects that disrupt the global network of atomic interactions and could thus impact distant regions of the protein. Our approach also has certain limitations. In particular, it does not account for the change in the conformation of the protein that could follow some of the mutations, as all the interactions, with the exception of the mutated residue, are kept identical. In addition, as our method aimed at measuring the impact of each residue individually as opposed to simulating the experimental setup, it is not stricktly speaking an alanine scanning method, since we removed all the side-chain interactions including those associated with the β carbon. Both issues could however be easily addressed if needed; the former through a short MD simulation, and the latter through a simple modification of the graph construction.

The strong agreement we obtained with the results from experimental binding assays was certainly not predictible. Our methodology is indeed not designed to capture changes in the binding free energy and its accuracy therefore suggests that the global structural organisation of the protein-ligand complex plays, at least in the case of MTIP, a major role in the binding strength. This observation calls for further studies to evaluate how the binding affinity in proteins is linked to their global structural organisation.

Several of the predictions presented in this chapter have been later confirmed by experiments (Douse *et al.*, 2012; Turley *et al.*, 2013) conducted after this study had been published (Delmotte *et al.*, 2011). In particular, the increased physical stability of the structures upon binding with MyoA, reflected by a decrease in the variability of the solutions in the variation of information, was verified by subsequent NMR experiments (Douse *et al.*, 2012), which identified a general loss of flexibility of the protein in the bound state. In the same article, Douse *et al.* (2012) showed

that residues from helix $\alpha 0$, which we identified as being exceptionally stronlgy disconnected from the rest of the N-terminal domain (see Figure 5.3), do not actually form a helix in solution but remain as an unstructured region which contributes neither to the binding with MyoA nor to the folding of the rest of the protein. In addition, they observed helix $\alpha 8$, the last helix to join the C-terminal domain in our analysis and therefore the most independent region according to our measure, to be in a different dynamic regime from the rest of the C-terminal domain. Finally, the hypothesis of the absence of an extended central α-helix in physiological conditions, which our results strongly supported, has been confirmed experimentally only very recently (Turley *et al.*, 2013).

Together, the results in this chapter provide a better understanding of the possible dynamical behavior of MTIP and other myosin light chains and their strong concordance with earlier as well as subsequent experiments confirms that our methodology is a good predictor for the structural organisation and dynamical behaviour of the protein.

Chapter 6

Highly multiscale biomolecular structures

In the preceding two chapters, we benchmarked our method on a well characterised protein example and demonstrated its predictive power on a recently discovered myosin light chain in the context of protein-protein interactions. In this chapter, we focus on large protein complexes displaying a highly multiscale structural organisation and covering a much broader range of length and time scales, and introduce new developments in our methodology to probe this increased complexity. In particular, we use Markov stability to explore phenomena taking place at the extreme end of the spectrum of scales, beyond the tertiary structure, such as the interplay between subunits of multimeric structures, or patterns of communication spanning the entire complex.

In the first part, we develop a detailed analysis of Rubisco, a vital yet notoriously inefficient enzyme responsible for carbon assimilation in photosynthetic organisms. A large heteromeric enzyme, with four dimers of large subunits and eight small subunits in its most common form, Rubisco exhibits highly complex multiscale dynamics associated with its multistep catalytic reaction. The structural mechanisms which control its activity, and particularly the fundamental roles of the quaternary structure interactions are still poorly understood. Using an ensemble of structures which capture the conformation of Rubisco at several key intermediate steps of its reaction, we explore the evolution of its structural organisation both through scales and throughout the catalytic pathway. In the second part, we contrast our results with two classic examples of multimeric structures representing extreme cases of a simple and a very elaborate hierarchical organisation: hemoglobin and ATCase.

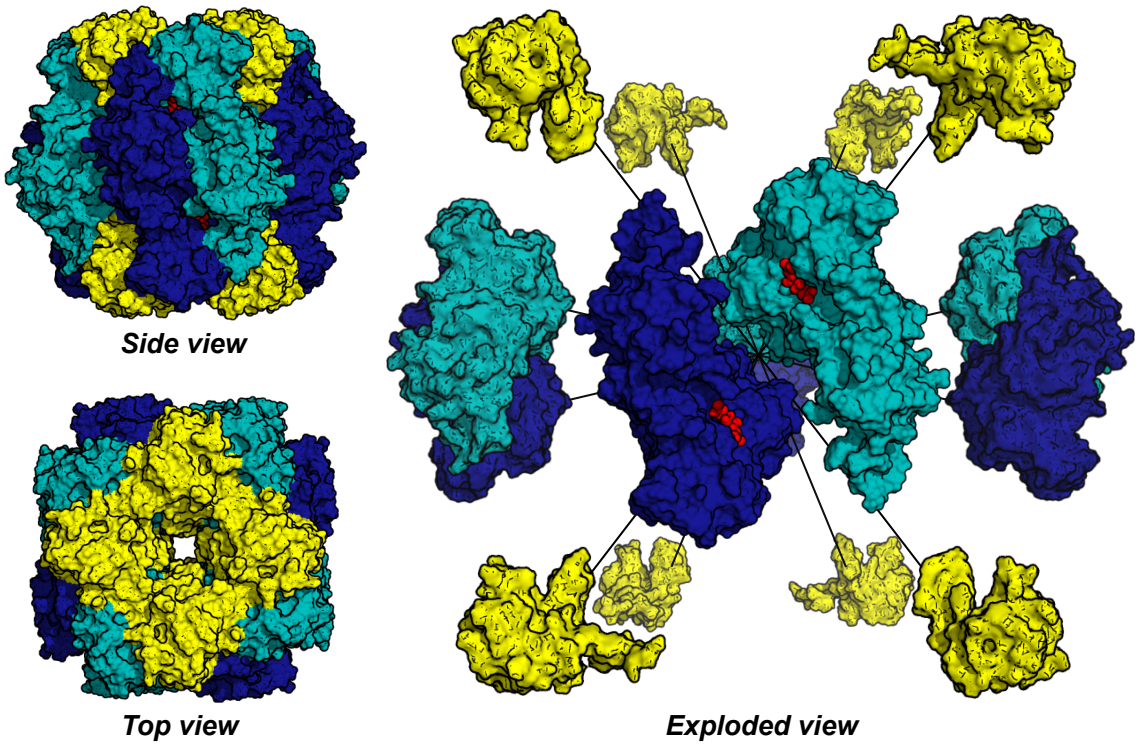


Figure 6.1: The structure of hexadecameric form I Rubisco contains a barrel formed by four pairs of large (L) subunit dimers which each enclose two active sites, and capped at both ends by four small (S) subunits. Small subunits are colored in yellow, large subunits in blue and cyan, and the substrate RuBP, located in the active sites, in red. Structures drawn with PyMOL using the PDB structure 1RXO.

6.1 The structure and function of Rubisco

6.1.1 Context and perspectives

Ribulose 1,5-bisphosphate carboxylase/oxygenase (Rubisco) is often considered to be one of the most important enzyme on the planet. It catalyses the fixation of atmospheric CO_2 during the Calvin cycle in a reaction that constitutes the source of virtually all the organic carbon in the biosphere, the building material for the organic molecules of life (Hartman and Harpel, 1994; Field, 1998; Schneider *et al.*, 1992). The carbon atoms in our own cells, our food, our clothes or the fuel we consume have probably all one day gone through the active site of Rubisco.

In spite of its vital role, Rubisco is notoriously inefficient with between only one and four carboxylation reactions completed per second at each catalytic site (Parry *et al.*, 2007; Schneider *et al.*, 1992). In addition to its slow rate, it suffers from a lack of specificity. Having first evolved three to four billion years ago in an atmosphere

much richer in CO_2 and depleted from oxygen (Kapralov *et al.*, 2011), it cannot completely distinguish between O_2 and CO_2 in the modern atmosphere. In addition to its carboxylation reaction, it catalyses a competing reaction with atmospheric oxygen which depletes the pools of substrate and generates a waste product that is eventually degraded through an energy consuming salvage pathway, resulting in a lowering of its efficiency by up to 50% (Ellis, 2010; Lorimer and Andrews, 1973; Zelitch, 1973). It has been hypothesised (Ellis, 2010) that, to account for its poor turnover, plants have multiplied their content in Rubisco. It is thought to be the most abundant protein on Earth, Rubisco forming an estimated 50% of the soluble proteins in plant leafs.

The impact of Rubisco's catalytic role on photosynthetic efficiency and crop yield is considerable and has established Rubisco as a potential optimisation target to address societal challenges such as the food and energy crises (Monteith and Moss, 1977). There are substantial variations between species in terms of both Rubisco's catalytic rate (Sage, 2002) and specificity (Bainbridge *et al.*, 1995; Jordan and Ogren, 1981, 1984; Parry *et al.*, 1989; Read and Tabita, 1994), and it was anticipated that structural differences could be exploited to guide genetic modification to improve Rubisco's efficiency. However, numerous attempts at mutant and chimeric variations of Rubisco have met with limited success, notably due to an inverse correlation between catalytic rate and specificity (Bainbridge *et al.*, 1995; Tcherkez *et al.*, 2006; Zhu *et al.*, 2004; Tabita *et al.*, 2008; Tabita, 1999; Spreitzer, 1999; Whitney *et al.*, 2011). Unravelling the engineering rules that control and regulate Rubisco's activity therefore appears to be critical if we are to reveal, and ultimately manipulate the regulatory restraints that dictate photosynthetic efficiency.

6.1.2 Complexity in Rubisco's structure and functional mechanisms

Uncovering the relationship between structural modifications and catalytic efficiency in Rubisco remains a major challenge due to, firstly, the remarkable intricacy of Rubisco's large heteromeric structure and, secondly, the complexity of the associated functional mechanisms (Portis Jr, 1992; Schneider *et al.*, 1992; Cleland *et al.*, 1998)

Rubisco is one of the largest enzymes in nature (550 kDa) and exists in a variety of multimeric forms. The most common form, which will be the focus of this work, is found in higher plants, cyano-bacteria and eukaryotic algae (Hartman and Harpel, 1994; Spreitzer, 1999; Spreitzer and Salvucci, 2002; Andersson, 2008). It is a heterohexadecamer (16 subunits) and comprises four dimers of large (L) sub-

units (50-55 kDa), each enclosing two active sites, forming a barrel capped at the top and bottom by four small (S) subunits (12-18 kDa) (see Figure 6.1). In some species of autotrophic bacteria and archaea, it can take the form of a homodimer, homooctamer or homodecamer of exclusively large subunits.

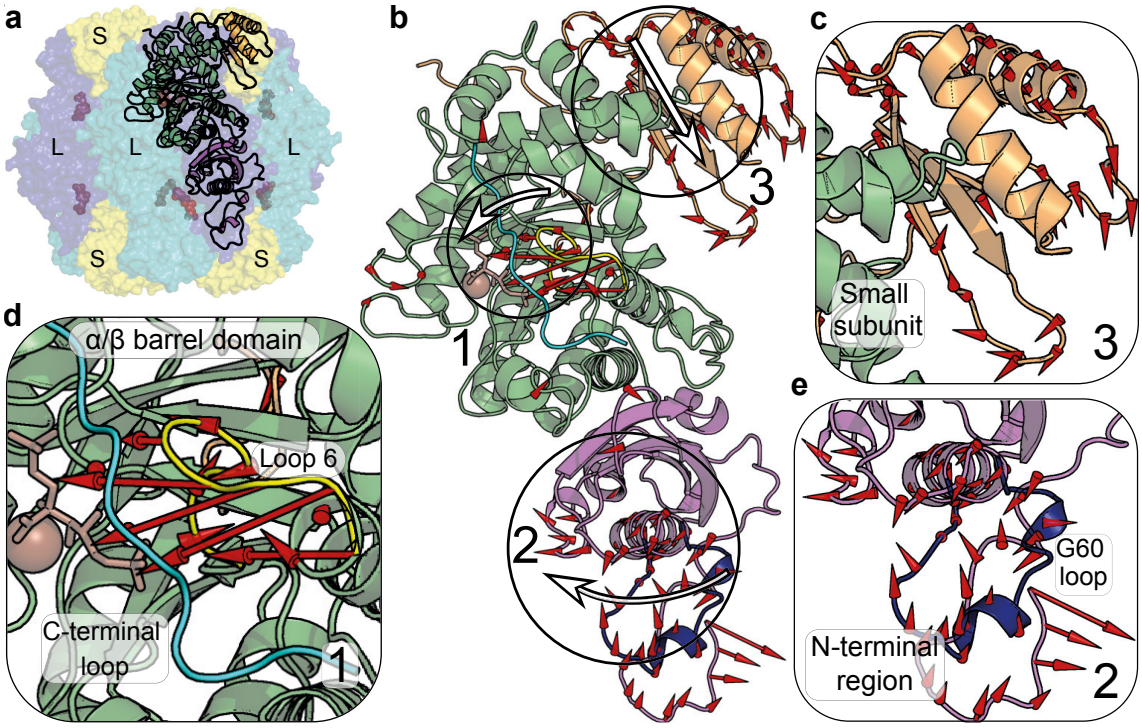


Figure 6.2: **a.** Hexadecameric Rubisco (1RXO) showing the location of large (L) and small (S) subunits. **b.** Conformational changes in Rubisco during closure involve motions over a wide range of scales, including loop motions, domain rotation and subunit displacement, here indicated by red arrows superimposed on one L and one S subunit. The conformational changes observed between open and closed forms are obtained by aligning the structures 1RXO (Stage II, open) and 8RUC (Stage III, closed) using PyMol with a cutoff of 0.6Å. The length of the arrows is equal to the displacement. **c, d & e.** Zoomed regions of **b.** indicating the motions observed during closure in the α/β barrel domain (**d**, in green), N-terminal domain (**e**, in purple), and S subunit (**c**, in orange). These motions involve the closing of loop 6 (**d**, in yellow) into the active site (ligand and magnesium ion colored in brown) located inside the barrel domain (**d**, in green), the rotation of the N-terminal domain (in purple and blue in **e**), the ordering of the C-terminal loop along the barrel domain (in cyan in **d**), and the displacement of the S subunit (in orange in **c**). Colors used in this figure reflect the organisation of Rubisco as suggested in the literature (Duff *et al.*, 2000; Taylor and Andersson, 1996), and are not the result of our analysis.

Whilst the large subunits, which contain the active sites, are directly involved in the reaction (Portis Jr, 1992; Hartman and Harpel, 1994; Cleland *et al.*, 1998; Spreitzer, 1999; Tabita, 1999; Spreitzer and Salvucci, 2002), the small subunit is not essential for catalysis and its precise role remains unclear. It is notably absent

in some active wild-type species (Tabita *et al.*, 2008) and an artificial structure of *Synechococcus* deprived of its small subunits has been shown to retain some activity, although as little as 0.6% of the turnover rate (Gutteridge, 1991). In addition, small subunits display a wide diversity in sequence and structure, associated with significant variations in stability, specificity and catalytic rate (Spreitzer and Salvucci, 2002; Spreitzer, 2003), as shown in chimeric enzymes combining large and small subunits from different species (Karkehabadi *et al.*, 2005).

This structural complexity is associated with a very elaborate functional mechanism which is not yet well understood. Rubisco's catalytic function involves a sequence of steps associated with structural rearrangements, which have been characterised through detailed comparisons of crystal structures that captured Rubisco's conformation at different stages of the reaction (Duff *et al.*, 2000; van Lun *et al.*, 2011) (see Figure 6.3 for a visual summary of the structural reaction intermediates, and Figure 6.2 for details of the conformational changes). First, Rubisco is activated through the carbamylation[1] of the Lys201 residue in the active site and subsequent stabilisation with Mg^{2+}. This step is followed by the binding of the substrate ribulose-1,5-bisphosphate (RuBP) which is linked to a conformational change from an 'open' to a 'closed' conformation and a series of reaction steps (enolisation, carboxylation and hydration reactions, see Figure 6.3), although the precise timing remains unclear. The open-closed conformational change involves the closure of loop 6 (in yellow in Figure 6.2d, motion 1), the rotation of the N-terminal domain (in purple in Figure 6.2e, motion 2), the displacement of the entire S subunit (in orange in Figure 6.2c, motion 3), and, finally, the ordering and packing of the C-terminal strand (in cyan in Figure 6.2d) along the barrel domain and against loop 6 (Duff *et al.*, 2000; Taylor and Andersson, 1996). The process ends with the opening of the enzyme and the release of two molecules of 3-phosphoglycerate (PGA). Rubisco can be inhibited by several molecules, including CABP, a transition state analog, and XuBP, which results from the isomerisation of the substrate RuBP. In addition, if RuBP binds before the active site is activated, Rubisco becomes inhibited.

The 'open' to 'closed' transition is thus associated with displacements within the individual L subunits. Such local rearrangements are embedded within a highly organised heteromeric quaternary structure supporting inter-subunit collective phenomena that could be linked to regulatory mechanisms such as cooperativity between active sites or the role of effectors (Taylor and Andersson, 1996; Yokota *et al.*, 1991;

[1] Addition of a CO_2 molecule.

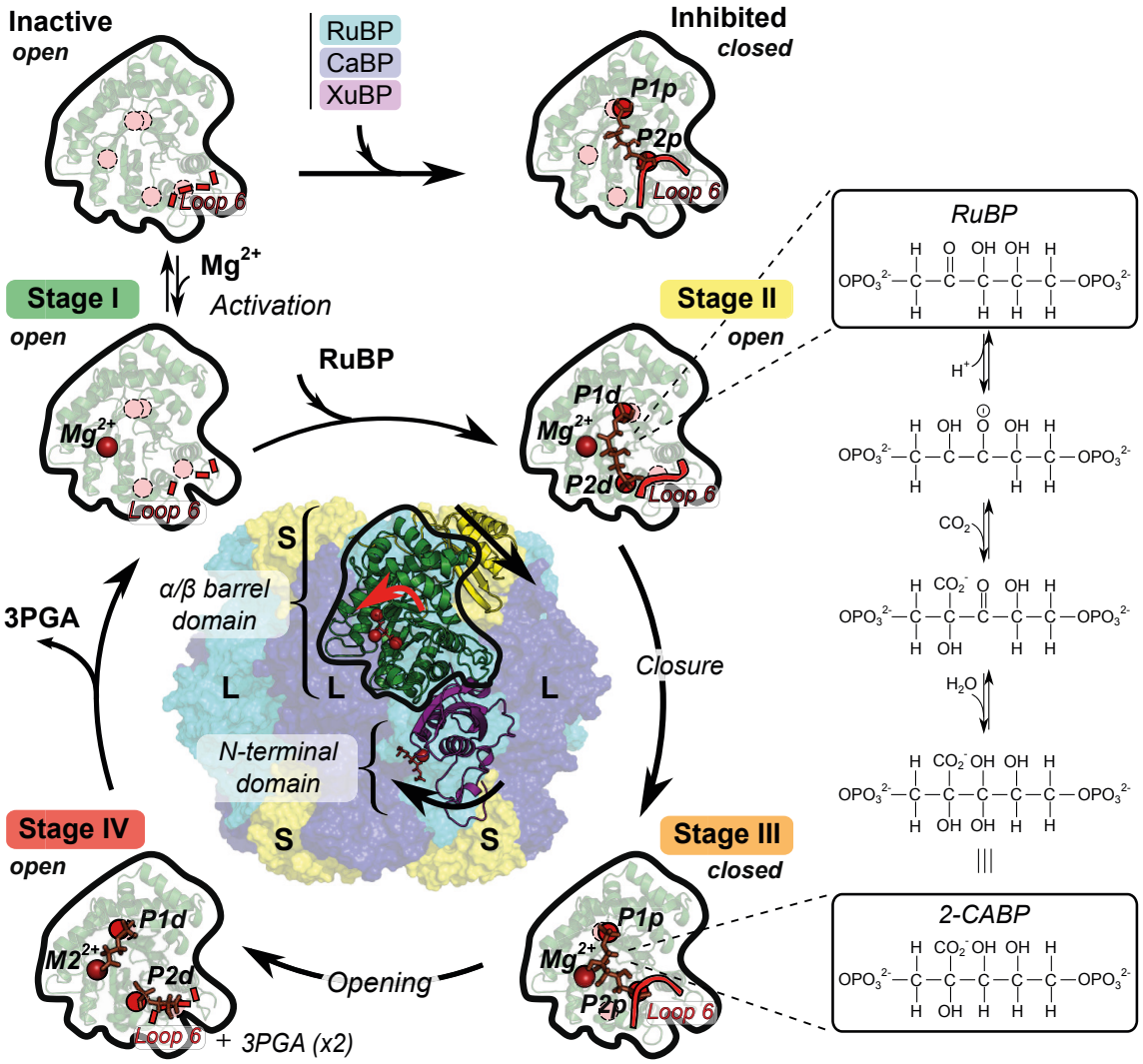


Figure 6.3: Rubisco's catalytic reaction steps and the corresponding structural conformations of the enzyme. Following the activation step (Stage I), the substrate ribulose-1,5-bisphosphate (RuBP) binds (Stage II) and undergoes enolisation, carboxylation and hydration reactions (reaction scheme shown leading to Stage III). Rubisco opens again at the end of the reaction (Stage IV), following the carbon-carbon cleavage and protonation reactions that give rise to the product 3-phosphoglycerate (PGA). The pictures at each stage indicate the motion of loop 6 and the ligand anchoring points within the α/β barrel domain (black line in the central inset). The anchoring points are colored red when a ligand is bound, and pink when free. Ligands such as xylulose 1,5-bisphosphate (XuBP) and 4-carboxy-arabinitol 1,5-bisphosphate (CABP) inhibit activity, as does RuBP when it binds to inactive Rubisco.

Parry *et al.*, 2008). However, the role of this larger scale of organisation, and in particular of the S subunits, remains unclear.

6.1.3 Computational means to address Rubisco's complexity

The computational exploration of Rubisco dynamics across different time and spatial scales can shed light on how its structural organisation affects the catalytic function. However, the vast array of length and time scales involved in the structural rearrangements associated with its catalytic function places the computational analysis of Rubisco beyond standard simulation methods. The use of fully atomic Molecular Dynamics simulations is severely limited in the case of Rubisco due to both its size and the slow time scales of the biologically relevant dynamics. On the other hand, coarse-graining techniques not only ignore atomic details but also decouple the different levels of organisation, and are thus unable to link atomic scale events, such as substrate binding, with the large-scale conformational changes induced.

Here we shed light on the link between chemical structure, conformational rearrangements and biological function using Markov stability. Our framework allows us to elucidate how the all-scale structural anatomy of Rubisco changes throughout the catalytic reaction keeping atomistic details throughout.

Our results uncover, for the first time to our knowledge, changes in the hierarchical organisation of Rubisco at *all scales* associated with different stages of the reaction and upon inhibition. In particular, at intermediate scales, we find that Rubisco's α/β barrel switches between the dominance of two hierarchical organisations, both coexisting and encoded as fingerprints in the structure. These hierarchies can be linked to structural rearrangements leading to an increased connectivity between the anchoring points of the ligand in agreement with experiments. At larger scales, we reveal differences at the level of the quaternary structure: the closure of the structure during the reaction leads to increased connectivity between the two lobes of the active site spanning across subunits. At even larger scales, we find a role for the enigmatic small subunits in mediating an enhanced connectivity between the large subunit dimers at the end of the catalytic reaction.

We also find evidence that the conformational changes that Rubisco undergoes during its enzymatic reaction or following inhibition are already encoded in the protein structure, similarly to the experimental suggestion that enzymatic behaviour is encoded in the intrinsic dynamics of the enzyme's unbound state (Henzler-Wildman and Kern, 2007). As our method accesses all scales without coarse-graining or re-

a Spinach Rubisco structures analyzed through Markov Stability

PDB ID	Conformation	Activation state	Ligand	Reaction stage
1AUS	***Open***	Activated	Mg^{++}	**I**
1RXO	***Open***	Activated	Ca^{++}/RuBP	**II**
8RUC	***Closed***	Activated	Mg^{++}/2-CABP	**III**
1AA1	***Open***	Activated	Mg^{++}/3PGA	**IV**
1RCO	***Closed***	Inactive	XuBP	**Inhibited**
1RCX	***Closed***	Inactive	RuBP	**Inhibited**
1RBO	***Closed***	Inactive	4-CABP	**Inhibited**

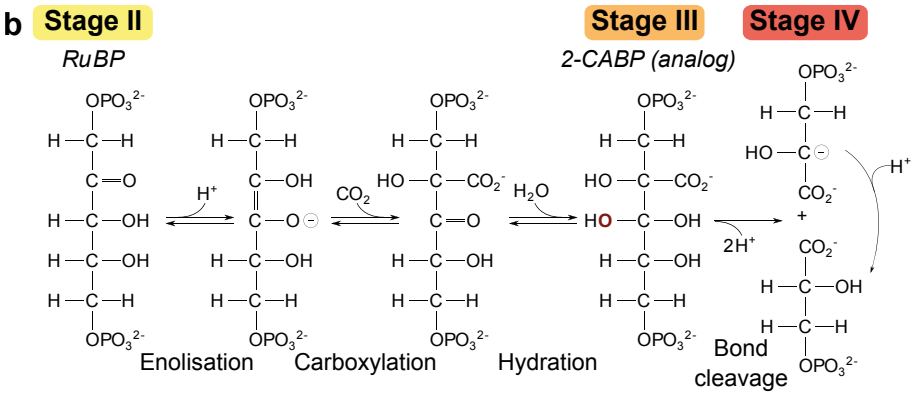


Figure 6.4: **a.** Details of the seven spinach Rubisco crystal structures used in this study: 1AUS (Taylor and Andersson, 1996), 1RXO (Taylor and Andersson, 1997b), 8RUC (Andersson, 1996), 1AA1 (Taylor and Andersson, 1997a), 1RCX (Taylor and Andersson, 1997b), 1RBO (Taylor *et al.*, 1996) and 1RCO (Taylor *et al.*, 1996). **b.** Rubisco's catalytic reaction steps associated with the corresponding structural conformations of the enzyme. The structure 8RUC used as a proxy for the transition state analog (Stage III) differs from the *in vivo* hydrated intermediate in the oxygen atom colored in red.

parameterisation, probing structure-function relationships throughout the catalytic cycle based on full atomistic detail, this may help identify distinct chemical design strategies for enzyme optimisation.

6.2 Materials and methods

6.2.1 Structural data

In this chapter, all the structures analysed are X-ray crystal structures of hexadecameric (L_8S_8) wild type forms of *Spinacia oleracea* (spinach) Rubisco. Spinach was chosen due to the availability of crystal structures associated with a series of stages of the catalytic reaction and bound to several inhibitors found *in vivo*. Figure 6.4

summarises the structures used in this work in the context of the catalytic reaction steps.

Relevance of the structures as intermediates in the catalytic pathway

Although many stages of the catalytic pathway are represented in our dataset, the structures used for stages II and III are only proxy for the real conformation. Indeed, the structure of Mg^{2+}-activated Rubisco bound to RuBP (Stage II) has never been resolved and its exact conformation *in vivo*, whether it is open or closed, thus remains unknown and a subject of debate. Initially, the larger size of Ca^{2+} compared to the true activator ion Mg^{2+} was thought to artificially maintain the active site in an open conformation (Duff *et al.*, 2000). However, a closed activated structure with Ca^{2+}/CABP was later resolved (Karkehabadi *et al.*, 2003), which argues against calcium alone preventing closure. In the absence of any further structural data for Rubisco with Mg^{2+}/RuBP, we have assumed the Ca^{2+}/RuBP structure (1RXO) to be a reasonable proxy for the Mg^{2+}/RuBP state found *in vivo*, and will consider it as the equivalent of Stage II in the rest of this work. Regarding Stage III, CABP is the transition state analog to this reaction stage, but differs from the true hydrated intermediate by one oxygen atom (in red in Figure 6.4b). The structure of spinach Rubisco with CABP is the closest available structure to the conformation found *in vivo* and is considered as its best proxy in this work.

Treatment of the missing residues and ligands

For four of the Rubisco structures we used (1AUS, 1RXO, 8RUC, 1AA1), only half of the hexadecamer was reported in the PDB file. The full structures were generated by symmetrisation using the script `MakeMultimer.py`[2]. For PDB structures 1AUS (Taylor and Andersson, 1996) and 1AA1 (Taylor and Andersson, 1997a), the unresolved residues 333-337 from loop 6 were added based on the PDB structure 1RXO (Taylor and Andersson, 1997b) in order to avoid an unrealistic break in the chain of covalent bonds, yet any new weak interaction formed by these added residues were omitted. All the ligands have here been included in the protein graphs. Charges were modelled using the server PRODRG (Schüttelkopf and van Aalten, 2004).

[2]Available at `http://watcut.uwaterloo.ca/makemultimer/`

Solvent accessible area

The total solvent accessible area is computed using PyMol (Schrödinger, LLC, 2010) with a 1.4Å probe and, for comparability, is restricted to the common subset of residues resolved in all structures, i.e. residues 20-332 and 338-463 for the L subunits, and residues 1-123 for the S subunits.

6.2.2 Markov stability analysis

Due to the computational cost of evaluating the matrix exponential for large networks (Rubisco structures contain more than 70 000 nodes), we use here the linearised version of Markov stability with the combinatorial Laplacian (see Chapter 3):

$$R(t) \approx (1-t)R(0) + tR(1) \tag{6.1}$$

For each Markov time, the Louvain optimisation is repeated for one thousand random initialisations and the ensemble of solutions found is kept. We then report the optimal of all the solutions found at each time and we also calculate the mean variation of information (VI) of the ensemble of solutions obtained.

As we observed in AdK and MTIP in the last two chapters, sudden drops in the variation of information often convey the identification of a robust partition or community. The "freezing" of a portion of the graph into a well-defined community manifests itself by a drop in the total variation of information as this region of the graph ceases to contribute to it. In order to avoid the cumbersome analysis of a large ensemble of surrogates and z-score, we will here use this property in conjunction with plateaus in the number of communities as the main indicators of a relevant community structure.

6.3 The all-scale analysis of Rubisco

The all-scale Markov stability analysis of the atomic graph of activated unliganded spinach Rubisco (PDB 1AUS) is shown in Figure 6.5a. The Markov zooming starts by detecting chemical groups at high resolution, then on to amino acids and secondary structures, followed by intra-unit functional domains (labelled a–d in Figure 6.5a), and finally groupings involving the small subunits and spanning the quaternary structure organisation (labelled e–g). From Markov time 10^6 onwards, Rubisco exhibits a marked hierarchy of well-defined communities indicated by local

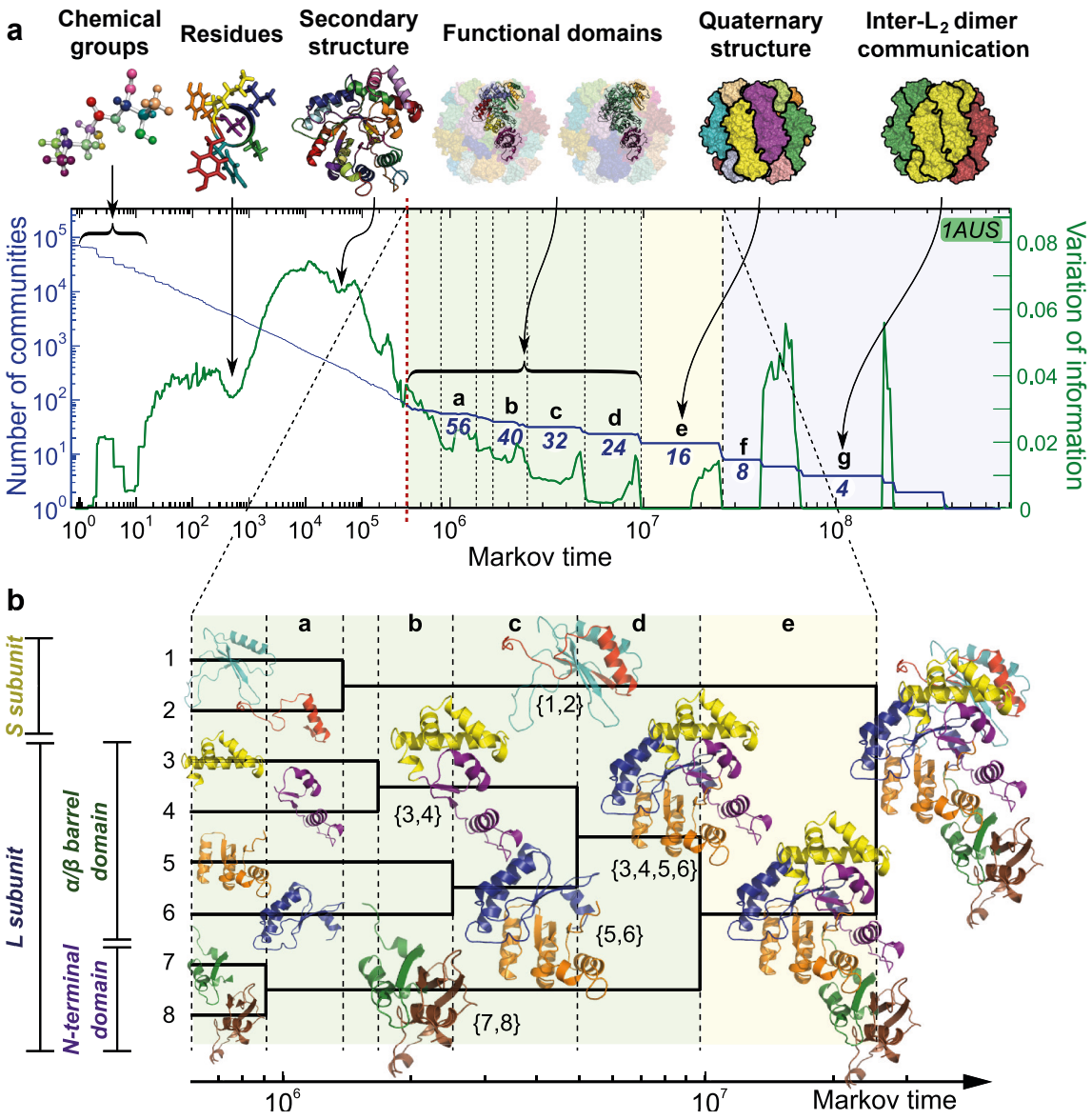


Figure 6.5: Structural anatomy of Rubisco at all scales. **a.** All-scale Markov stability analysis of activated unliganded spinach Rubisco (Stage I, PDB code 1AUS). As the Markov time increases, we recover first the meaningful chemical and biochemical levels of organisation (chemical groups, residues, secondary structure); then previously reported functional domains (N-terminal and barrel domains); and finally partitions involving the quaternary structure (L and S subunits) and inter-dimer interactions. The number of communities (blue line) decreases with Markov time indicating coarser partitions. Relevant partitions are indicated by persistent plateaus of the number of communities (blue line) together with dips in the variation of information (green line). **b.** Expanded view of the hierarchy of intra-unit functional domains. The N-terminal domain $\{7, 8\}$ and the S subunit $\{1, 2\}$ form robust communities persistent over a long range of Markov times.

minima in the variation of information (i.e. robust to the optimisation) and by long plateaus in the number of communities (i.e. persistent under the diffusive dynamics), as described in the previous chapters. However, in contrast to the results we obtained for AdK and MTIP, Rubisco displays a much larger number of robust community structures beyond the identification of the secondary structure. These communities are not only more numerous, but also span a much broader range of Markov times, and are associated with longer plateaus and smaller values of the variation of information. These features reflect the highly organised multiscale structure of Rubisco and distinguish it from proteins with less elaborate dynamics, as we will discuss in the second part of this chapter.

We have applied this method to study how the all-scale structural anatomy of Rubisco changes between the conformations it adopts throughout its catalytic reaction. Specifically, we have analysed four spinach Rubisco crystal structures, which provide snapshots of Rubisco at different stages of the reaction, labelled in this work from I to IV: activated unliganded (1AUS, Stage I); bound to the substrate RuBP (1RXO, Stage II); the intermediate state analog (8RUC, Stage III); and bound to the product 3-phosphoglycerate (PGA) (1AA1, Stage IV). We have also compared them with three conformations of inhibited Rubisco bound to three different ligands: xylulose 1,5-bisphosphate (XuBP) (1RCO), RuBP (1RCX) and 4-carboxy-arabinitol 1,5-bisphosphate (CABP) (1RBO) (see Figure 6.4).

Our analysis follows Rubisco's hierarchical organisation bottom up from its atomic organisation. As expected, our method finds no differences between the structures at short scales, since the chemical and biochemical building blocks are identical and the secondary substructures are largely similar across all conformations. However, at intermediate to long scales our analysis reveals significant differences in the structural organisation of Rubisco throughout the reaction and upon inhibition, involving the intra-unit organisation of the functional regions (α/β barrel, N-terminal domain) within the single L subunits, and the quaternary structure organisation (involving inter-subunit communication) at larger scales, as we now detail.

6.3.1 Intermediate scales: The intra-unit functional domains throughout the catalytic reaction

S subunits and N-terminal domain are decoupled from the barrel domain at intermediate scales

At intermediate scales, our analysis extracts the hierarchical organisation of intra-unit substructures, as seen in the expanded view in Figure 6.5b. Reassuringly, we find robust communities corresponding to the three functional domains in the enzyme (Taylor and Andersson, 1996): the L subunit α/β barrel domain $\{3, 4, 5, 6\}$, the L subunit N-terminal domain $\{7, 8\}$, the entire L subunit $\{2$–$8\}$ and the S subunit $\{1, 2\}$.

As shown in Figure 6.5b, both the S subunit $\{1, 2\}$ and the N-terminal domain $\{7, 8\}$ are particularly well conserved across scales (as shown by a very long branch in the hierarchical tree), and remain well separated from the α/β barrel across these intermediate Markov times. The independence of the S subunit is expected, due to the absence of covalent bonds with the L subunit. Yet we showed in the case of MTIP (Figure 5.3) that robust communities grouping sections from different subunits can form under the Markov stability optimisation, and the absence of such communities in Rubisco suggests that the dynamics of the small subunits is largely decoupled from that of the large subunits at intermediate scales.

This conclusion is supported by the fact that the S subunit is known not to be essential for catalytic function. However, it does not necessarily contradict reports of experiments observing a strong decrease in activity upon the removal of the S subunit from hexadecameric Rubisco. As it will be shown hereafter, our calculations suggest that S subunits play an important role at high scales.

At first sight, the isolation of the N-terminal domain $\{7, 8\}$ from the catalytically important α/β barrel domain $\{3$–$6\}$ of its own subunit is more surprising, as they are covalently linked. However, in this case, Markov stability detects sensitive details of the quaternary structure related to function. Hexadecameric Rubisco is indeed organised into four functional L subunit dimers (see Figure 6.1). The two L subunits forming a dimer are arranged antisymmetrically such that each dimer possesses two active sites enclosed by the α/β barrel domain $\{3$–$6\}$ of one L subunit together with the N-terminal domain $\{7, 8\}$ of the other. This result thus shows the capacity of Markov stability to uncover, in the case of Rubisco, the functional organisation into active sites beyond the apparent structural organisation of the hexadecamer.

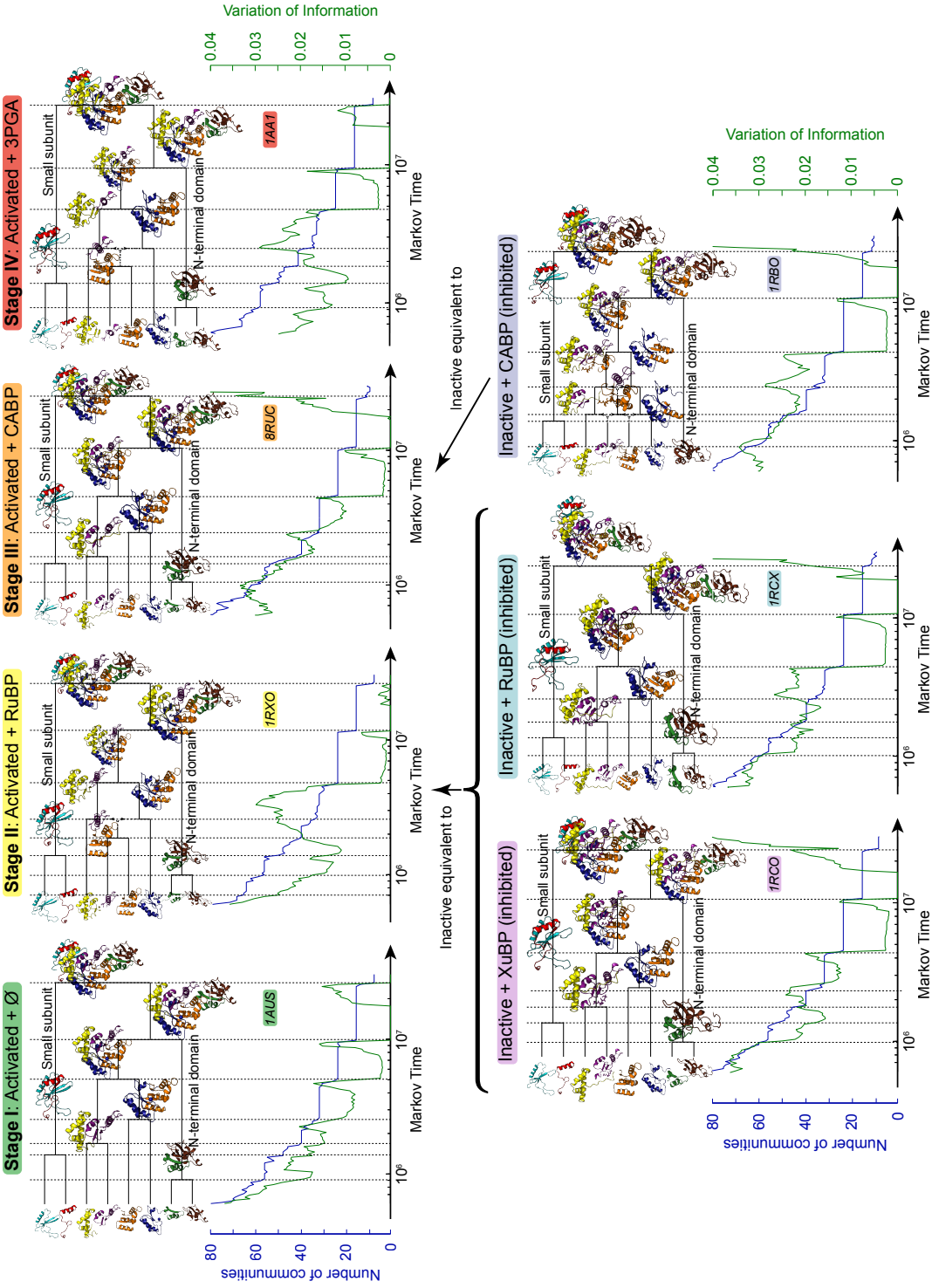


Figure 6.6: (Figure caption on the next page)

Figure 6.6: (Figure on the previous page) Markov stability analysis of the seven spinach Rubisco structures used in this study (see Figre 6.4). Each subfigure shows the quasi-hierarchical organisation identified by Markov stability for each PDB structure, together with the evolution of the total number of communities (in blue) and variation of information (in green) as a function of the Markov time. All spinach Rubisco structures share a number of similarities in their structural organisation, including the presence of well-defined communities for the N-terminal domain and S subunit. The analysis of inactivated structures (1RCX, 1RCO and 1RBO) suggests that, at intermediate scales, the absence of activation induces the opposite behaviour in the α/β barrel hierarchical organisation to the equivalent activated structure (see Figure 6.8).

Generalities across structures

The four communities identified in the previous section relate to fundamental elements of the functioning and structure of Rubisco (Taylor and Andersson, 1996; Duff *et al.*, 2000), and should thus be well conserved in all conformations. Figure 6.6 shows the results of the Markov stability analysis for all seven structures: As expected, these four important regions of the protein, i.e. the N-terminal domain, α/β barrel domain, L and S subunits, indeed form well defined communities in all the structures of spinach Rubisco we analysed.

Similarly, our previous conclusion that the small subunits are largely decoupled from the large subunits at intermediate scales is general and can also be observed in the other six structures. To verify this hypothesis, we carried out the Markov stability analysis of all structures with the S subunits removed from their protein graph, thus effectively creating their equivalent L_8 structure. Figure 6.7 confirms both L_8 and full L_8S_8 structures yield identical results up to the Markov time at which the quaternary structure is identified. At Markov times beyond 10^7, the curves start to diverge at the scale where the small subunits begin to influence the quaternary structure organisation, which we will explore in more details later. These results thus support the hypothesis of a weak involvement of the S subunits in the dynamics of the L subunit at intermediate scales at all stages of the catalysis.

Evolution of Rubisco structural organisation during catalysis

Now that we have identified these four well conserved regions, we focus on the changing elements of Rubisco's structural organisation throughout the catalytic pathway.

As it is shown of Figure 6.6, the majority of the differences between the seven structures of Rubisco are localised inside the α/β barrel domain, the region containing most of the active site. The activated structures at Stages I and III show a

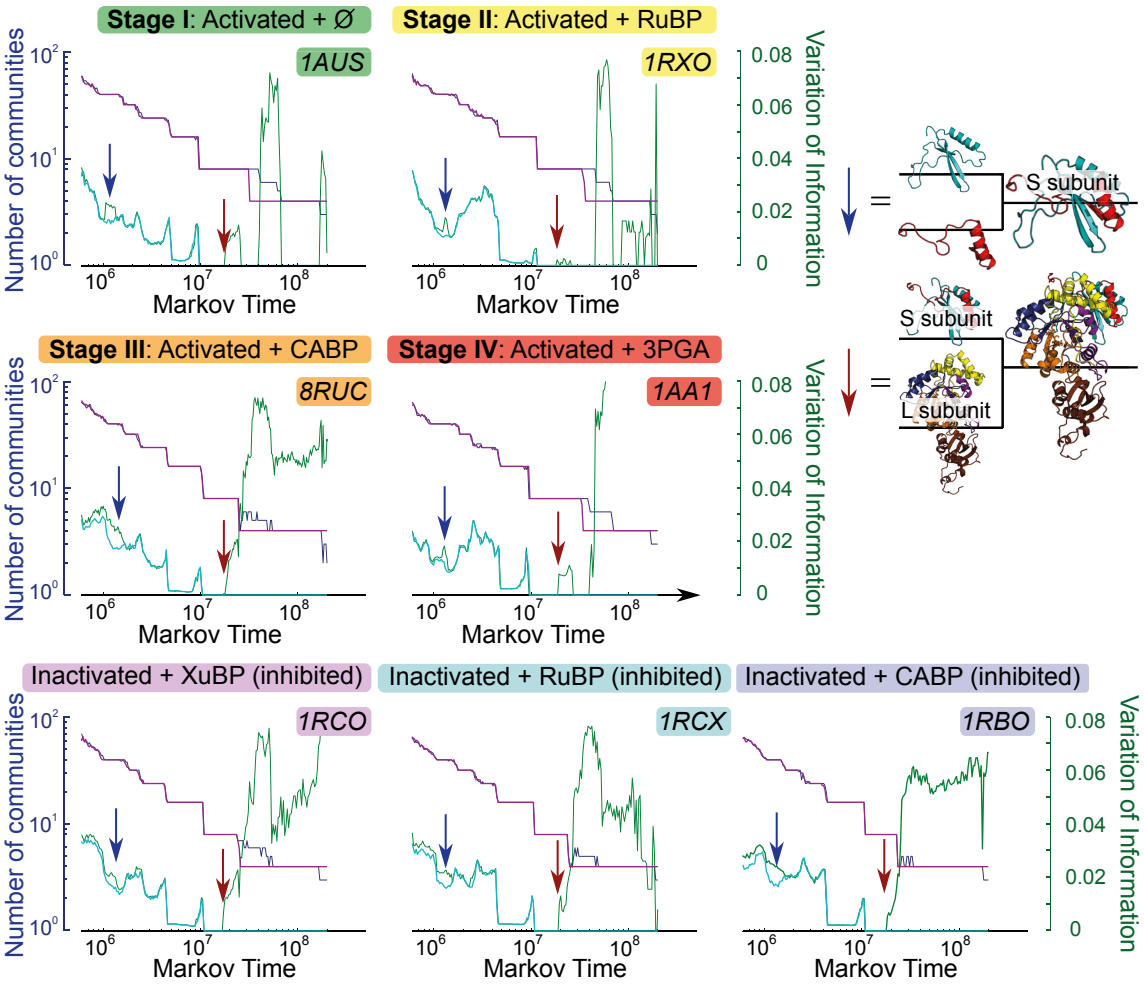


Figure 6.7: Comparison between the Markov stability analyses of the original L_8S_8 Rubisco structures (blue and green), and their L_8 counterpart (magenta and cyan) obtained by removing the small subunits from the original PDB structures. Up to the Markov time at which intersubunit groupings are identified (marked by the red arrow), the results of the L_8 and L_8S_8 structures are superimposed with the exception of internal rearrangements within the S subunits (marked by the blue arrow). Beyond the quaternary structure (beyond the red arrow), the Markov stability analyses diverge as the S subunits are then involved in the inter-L subunits communication.

robust structural hierarchy clearly subdividing the barrel domain into two and four communities. Stages II and IV, on the other hand, exhibit a less compartmentalised barrel region with a non-hierarchical organisation comprising two, three and four communities. These two hierarchies, labelled type 1 and 2 respectively, are shown in more details in Figure 6.8a.

These two types of hierarchies appear unrelated to the global (open/closed) conformation, which is often used to classify Rubisco structures. Whereas Stages I, II and IV are open and Stage III closed, type 1 hierarchy is common to Stages I and III and type 2, to Stages II and IV. (see Figure 6.3).

To resolve this apparent discrepancy, we analyse the ensemble of solutions, optimal as well as suboptimal, given by the Louvain algorithm at each Markov time, rather than focusing on the highest Markov stability partition only as in Figures 6.5 and 6.6. In some cases indeed, the ensemble of Louvain-optimised partitions can be dominated, not by a single solution, but several solutions with similar values of Markov stability (one optimal, and a few suboptimal but optimised). The Louvain ensemble of solutions therefore provides a more complete picture which include other highly relevant partitions, close in Markov stability to the optimal one.

To visualise the ensemble of partitions found and their relevance at each Markov time, we construct heatmaps of community structures. The heatmaps are computed from the observed frequency of each similar community in the ensemble obtained from 1000 runs of the Louvain optimisation at each Markov time. Communities are considered similar if: (i) they differ in fewer than 100 nodes ($\sim$ 6 amino acids) so as to allow for flexibility in communities with soft borders such as loops, which do not influence much the overall Markov stability of the graph; or (ii) they contain the same elements of the secondary substructure (even if they differ by more than 100 nodes), so as to include communities with longer soft borders.

In Figure 6.8, the heatmaps clearly highlight the presence of additional coexisting community structures besides the optimal partition. Each line of a heatmap corresponds to one particular community, and each column to a particular Markov time. The colours indicate the frequency with which each community is observed in the Louvain ensemble of optimal and suboptimal partitions, going from dark blue (not observed) to dark red (identified in 100% of the Louvain runs at this Markov time). The black lines superimposed on the heatmaps indicate the optimal Markov stability partition (which may differ from the most frequent), as obtained in Figure 6.5b.

We find that the two structural hierarchies identified previously (Hierarchies 1 and 2, Figure 6.8a) are both present at all stages, and are each alternatively more

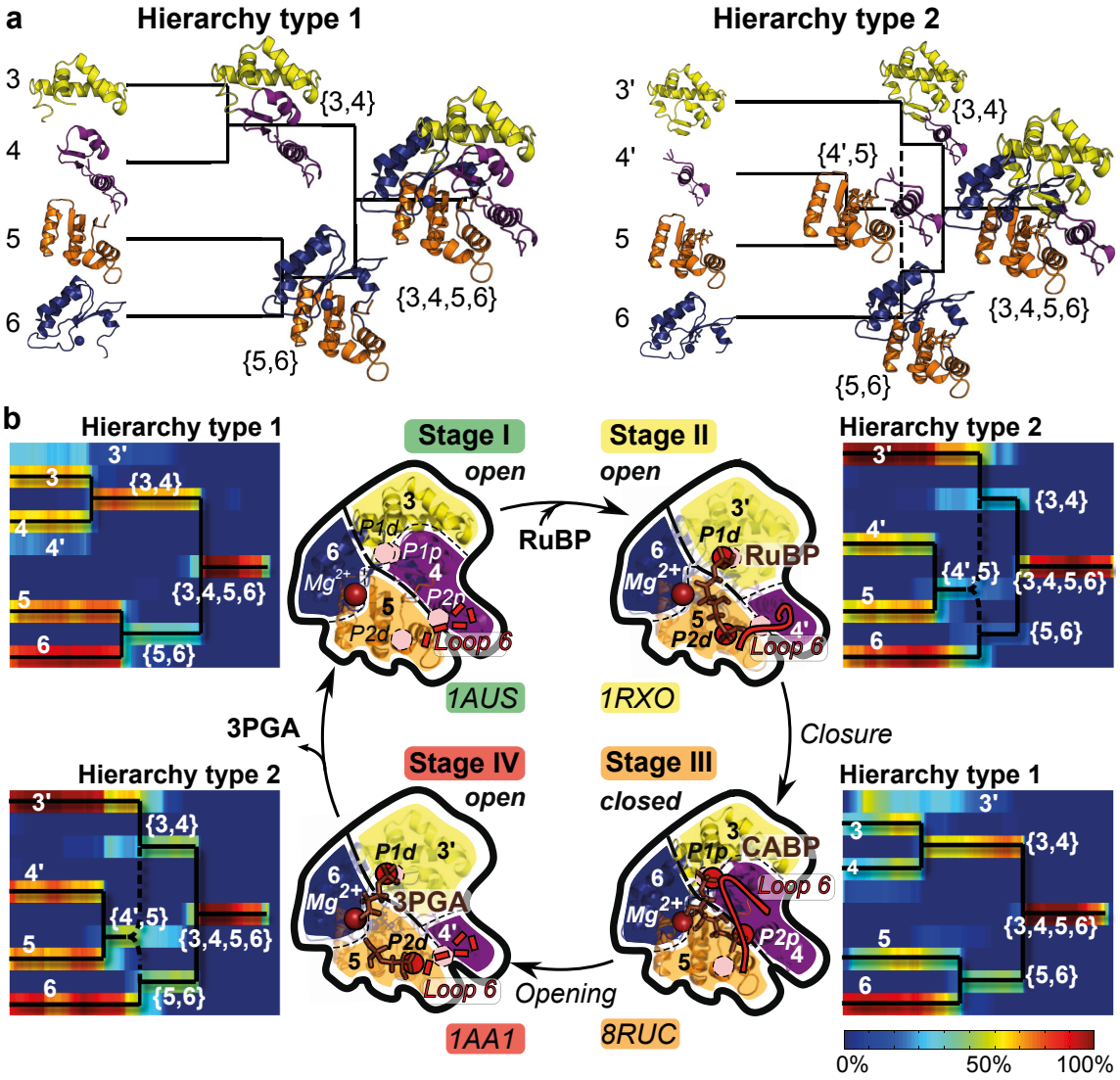


Figure 6.8: Rubisco switches between two hierarchical organisations of the α/β barrel throughout the catalytic reaction. **a.** The two types of hierarchical organisation of the barrel found with Markov stability in all seven Rubisco structures analysed as in Figure 6.5a. **b.** The heatmaps depict the probability of finding each community when Markov stability is optimised at each Markov time and show the coexistence of both hierarchies throughout the stages of the reaction. However, Stages I and III are characterised by the dominance of the well-defined Hierarchy 1, while Stages II and IV are characterised by dominance of Hierarchy 2, which has more diffuse communities encompassing regions around loop 6 and the two anchoring subsites of each phosphate group. The anchoring points are colored red when a ligand is bound, and pink when free.

probable: Stages I and III are both dominated by Hierarchy 1, with its well-defined organisation of the α/β barrel domain into two binary splits, and Stages II and IV are characterised by a more fluid structure of the barrel domain, as indicated by the less robust partitions of Hierarchy 2. The heatmaps however indicate that the suboptimal partitions of Hierarchy 1 are found during Stages II and IV, when Hierarchy 2 is optimal and, conversely, signatures of Hierarchy 2 are visible in Stages I and III, when Hierarchy 1 is dominant.

These results relate to previous observations from the preceding two chapters, but take place here in the context of an ensemble of possible states as opposed to a single one previously. Firstly, our results show that the hierarchy that dominates in the 'closed' transition-state conformation (Stage III) is already encoded in the 'open' unliganded structure (Stage I), similarly to our results for AdK and MTIP in the open and closed forms. Secondly, the presence of both hierarchies at all stages of the reaction, alternatively as optimal and suboptimal, is suggestive of the coexistence of two local minima in the landscape of protein conformations throughout the catalytic reaction encoded in the structure: the conformational changes associated with the binding of the substrate RuBP or the formation of the product PGA would thus shift the likelihood of these two hierarchies of intra-unit functional domains. Thirdly, the hierarchies obtained by our method are also consistent with other experimental observations from Rubisco crystal structures linking key functional aspects with rearrangements of loop 6 and the three anchoring points of the ligands during the catalytic reaction (Duff *et al.*, 2000). Previous comparisons of Rubisco crystal structures (Taylor and Andersson, 1996) have established loop 6 as a key structural element for the open-closed transition, and as the region of Rubisco undergoing the largest displacement between these main conformations. In addition, Duff *et al.* (2000) proposed a model linking the three ligand anchoring points to Rubisco's function and associated closing mechanism. The three anchoring points located in the active site (represented by red dots in Fig. 6.8) are the magnesium ion and the pockets formed around each of the two phosphate groups of the ligand (P1 and P2). During the transition from the open to the closed conformation (Stage II to III), the P2 site moves from the upper subsite (P2d, formed by Arg295 and His298) to the lower subsite (P2p, formed by Arg295 and His327), in a displacement which bends the ligand, reduces the P1-P2 distance, and removes its steric hindrance with loop 6, thereby allowing loop 6 to close into the active site. Similarly, the displacement of the P1 site from the distal subsite (P1d, Gly381, formed by Gly403, Gly404 and Trp66) to the proximal subsite (P1p, formed by Gly381, Gly403, Gly404 and Thr65)

is thought to trigger the rotation of the N-terminal domain of the neighbouring L subunit (see Figure 6.3).

Our analysis shows that Rubisco goes from Hierarchy 1 at Stage I, in which the residues forming the P1 anchoring points (both P1d and P1p) are spread out in different communities ($\{3\}$ and $\{4\}$) and P2d (in $\{5\}$) is isolated from P2p (spread across $\{4\}$ and $\{5\}$), to Hierarchy 2 at Stage II with fluid communities that, firstly, group together the P1 residues (in $\{3'\}$) and, secondly, the two subsites P2d and P2p ($\{4', 5\}$). Hence, Stage II reflects enhanced communication around the phosphate group anchoring points and between their subsites. Furthermore, the emergent role of loop 6 is signalled by its becoming part of a single community {4',5} at Stage II. Stage II can thus be viewed as a state with structurally-enhanced connectivity between the phosphate subsites, and across loop 6 and its surrounding substructures, hence primed for the closure of loop 6 needed for function (Duff *et al.*, 2000). Rubisco then recovers the Type I hierarchy at Stage III, but now with increased connectivity between the relevant functional points in the structure (loop 6 and anchoring points).

Finally, it is interesting to note on Figure 6.6, that inactive structures appear to have the opposite behaviour to their activated counterparts, with the inactivated forms bound to RuBP and XuBP (inactivated analog to Stage II) displaying a well defined structural hierarchy of type 1 and the inactivated structure with CABP (inactivated analog to Stage III) showing a fluid organisation similar to a hierarchy of type 2.

6.3.2 Large scales: closure favours inter-unit communication between the two domains of the active site

At larger scales, our method uncovers structural partitions beyond the individual subunits. This is significant for Rubisco since the active sites lie at the interface between two L-subunits. At these scales (*(e)* in Figure 6.5), our method always finds the expected optimal partition into 16 communities, i.e. the 'quaternary structure partition' with each of the sixteen L and S subunits as independent communities. However, coexisting with this natural partition, we find another sub-optimal 16-partition with communities that span across adjacent L-subunits enclosing each of the active sites. We denote this partition, which reflects the organisation of Rubisco in terms of catalytically competent substructures, as the 'catalytic region partition' (Figure 6.9a).

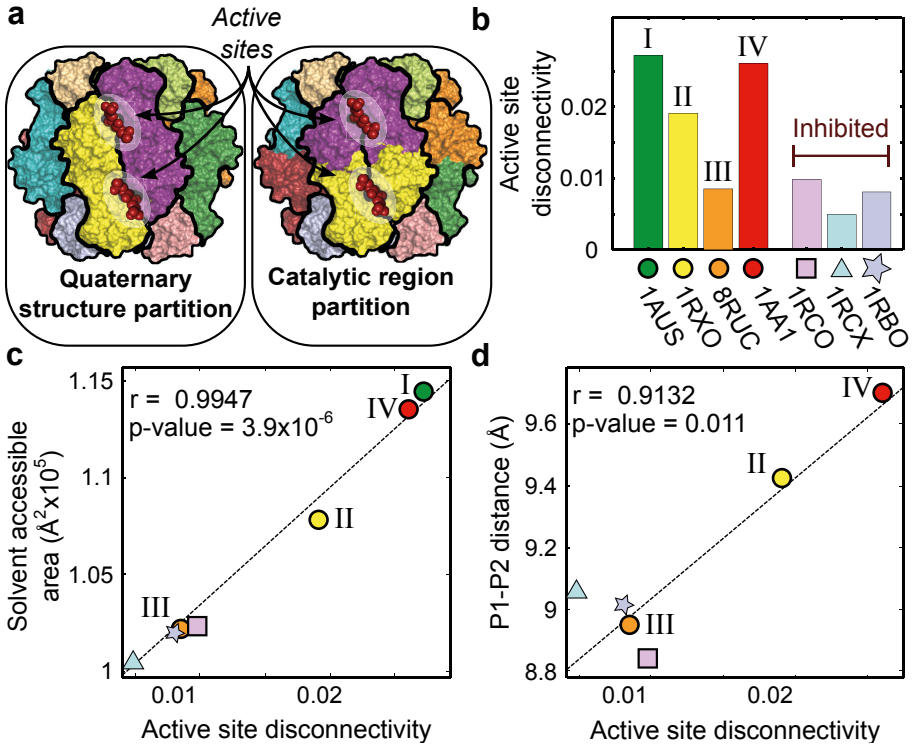


Figure 6.9: The disconnectivity between the two lobes of each active site is directly related to closure of Rubisco. **a.** The two 16-community partitions found: the optimal 'quaternary structure partition' and the sub-optimal 'catalytic region partition' spanning across units. **b.** The active site disconnectivity, defined as the difference between the Markov stability of the quaternary structure partition and the catalytic region partition, for all spinach Rubisco structures studied. **c.** Solvent accessible area and **d.** inter-phosphate distance (P1-P2) are both strongly correlated with the active site disconnectivity calculated in our analysis.

The difference in Markov stability between the optimal 'quaternary structure partition' and the suboptimal 'catalytic region partition' can be used as a measure of *disconnectivity* between the lobes of the active site situated in different subunits (α/β barrel of one L subunit and N-terminal domain of its partnering L subunit). Figure 6.9b shows that the disconnectivity is high for Stages I, II and IV, while the inter-L subunit communication across the two lobes of the active site is stronger at the transition state (Stage III) and in all the inhibited structures.

Furthermore, Rubisco undergoes a gradual transition between a fully open (Stage I) and a completely closed conformation (Stage III) in the course of the catalytic reaction. Closure of Rubisco's active site involves a rotation of the N-terminal domain towards the α/β barrel of the partnering L subunit, resulting in the formation of additional contacts between the two neighbouring L subunits which share the active sites (Taylor and Andersson, 1996). In Figure 6.9c & d we show that the active site

disconnectivity obtained with our method is closely related to two measures of closure calculated directly from the PDB structures, i.e. the solvent accessible area and the inter-phosphate P1-P2 distance previously suggested to regulate closure (Duff *et al.*, 2000).

Hence our results indicate that the additional interactions resulting from the rotation of the N-terminal domain towards the α/β barrel of the neighbouring L subunit during the 'open' to 'closed' conformational change induce increased communication across the active site spanning the subunits.

6.3.3 Ultimate scales: the role of the small subunits in enhancing connectivity across L_2 dimers

At even larger Markov times, our method uncovers graph communities at the highest level of the quaternary structure reflecting communication across L_2 dimers via the S subunits.

Were Rubisco to behave as four independent pairs of active sites, we would expect the optimal partition at long Markov times to have four well defined communities, one for each L_2 dimer enclosing a pair of active sites grouped with their corresponding S subunits (i.e. the S subunit with which each L subunit has the highest interaction energy). This partition corresponds to the L_2S_2 'tetrameric partition' in Figure 6.11a and is indeed the optimal partition for Stages I and II. However, the optimal partition at Stages III (closed) and IV (open), as well as in all inhibited structures (closed), is a 'promiscuous partition', in which each L_2 dimer swaps one of its S subunits with that of a neighbouring dimer—its two symmetric realisations are shown in Figure 6.11a.

The prevalence of the 'promiscuous partition' can be traced back to variations in the number and energy of hydrogen bonds across the three types of interfaces between the large and small subunits (van Lun *et al.*, 2011) throughout the catalytic reaction (see Figure 6.10). In the case of the tetrameric partition, the grouping between large and small subunits takes place across the highest energy L-S interface (LS1). The communities in the promiscuous partitions, on the other hand, involve all three L-S interfaces, both the high energy LS1 and the lower energy interfaces LS2 and LS3 which link the L_2 dimer to the S subunit 'belonging' to the neighbouring dimer. We find that the appearance of the promiscuous partition at Stages III and IV (and in all inhibited structures) is linked to an increased symmetry in the strength of the interaction energies of the LS1 vs. LS2/LS3 interfaces. While the total H-

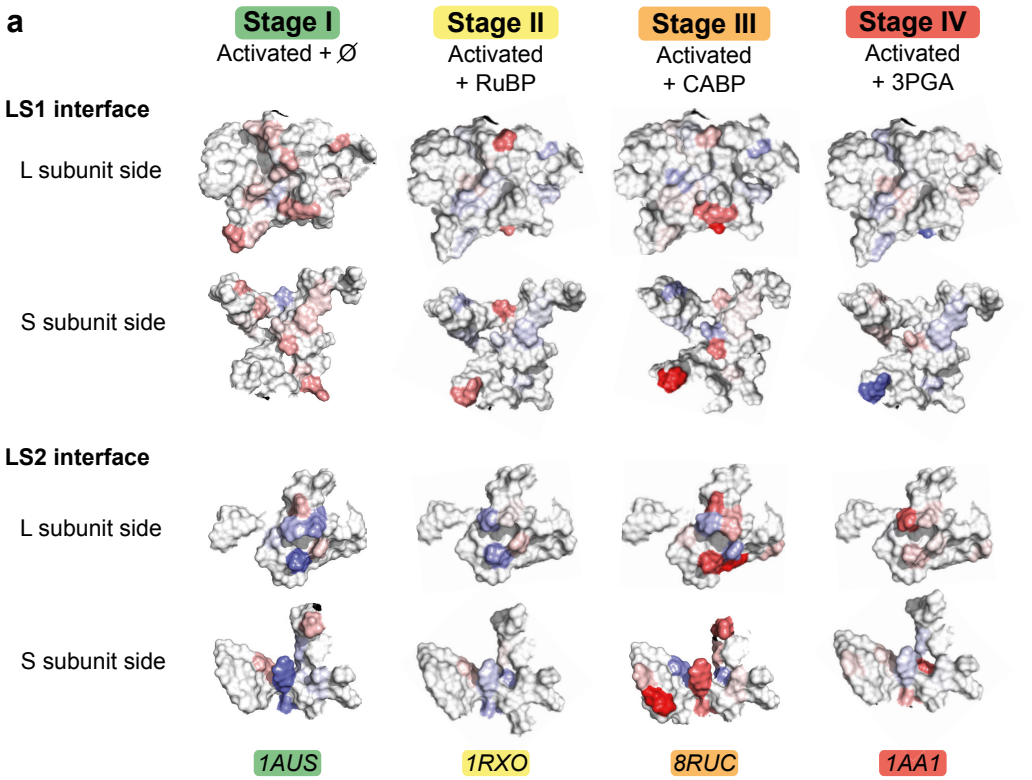
bonding energy across the LS1 interface clearly dominates that across LS2 and LS3 interfaces during Stages I and II, the interaction becomes almost balanced in Stages III and IV. The emergence of this symmetry effectively eliminates the preference of a S subunit for a particular L2 dimer and leads to the 'delocalisation' of the S subunit in the promiscuous partition with the result of increased connectivity in the overall structure. Interestingly, this symmetry persists when the structure 'opens' again in Stage IV, but is associated with a decrease in the total magnitude of the L-S interaction energy.

The difference in Markov stability between the 'tetrameric' and 'promiscuous' partitions can be used as an indicator of the inter-L_2 dimer connectivity via the S subunits. Indeed, because the 'promiscuous partition' has two symmetric, resonant realisations characterised by the 'flipping' of S subunits between adjacent L_2 dimers, it effectively creates a channel of communication between dimers through the 'delocalisation' of the shared S subunits. Our results, which we show on Figure 6.11b, thus suggest that Stages III and IV, as well as all inhibited structures are characterised by an increased communication between the catalytically competent L_2 dimers across the whole hexadecamer.

The efficiency of this inter-dimer communication channel mediated by the S subunits can also be estimated from the probability of finding each type of partition in the ensemble of solutions given by the Louvain optimisation, similarly to the heatmaps from Figure 6.8. The sharing of the small subunits between L_2 dimer should indeed manifest itself by a more diverse Louvain ensemble of partitions which would reflect that small subunits are being grouped indifferently with both L_2 dimers by the partitioning algorithm. This can be estimated using an information-theoretic measure, similar to a previously proposed estimator for the residue communication ability (Chennubhotla and Bahar, 2006), as

$$S_i = \sum_{j \in L_2} p_{ij} \log(p_{ij}), \tag{6.2}$$

where the sum extends over the L_2 dimers and p_{ij} is the ratio of partitions in which the S subunit i is grouped with the L_2 dimer j in an ensemble of 1000 Louvain optimisations of 4-community partitions between Markov times 9×10^7 and 1.5×10^8. Using this additional and independent measure provided by our method, we further confirm on Figure 6.11c the increased inter-L_2 dimer communication efficiency present in Stages III and IV (and in all inhibited structures) mediated via the shared S subunits.



b

	Stage I	Stage II	Stage III	Stage IV
LS1 (kJ/mol)	454.1	460.9	469.5	431.8
LS2 (kJ/mol)	337.3	318.8	405.9	353.4
LS3 (kJ/mol)	41.0	73.5	71.6	60.9
LS1+LS2+LS3 (kJ/mol)	832.4	853.2	947.0	846.2
Assymetry (LS1-(LS2+LS3))/LS1	0.17	0.15	-0.02	0.04

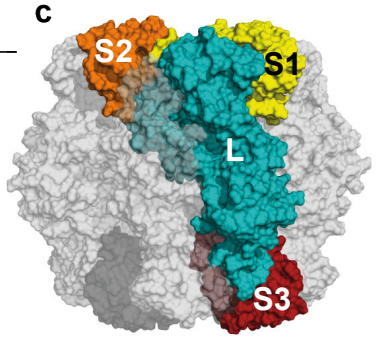


Figure 6.10: **a.** Areas of interface between the L and S subunits coloured as a function of the energy of hydrogen bonds (relative to the average): from blue (smaller than average H-bonding energy) to red (higher than average) through white (average). **b.** Energy of interaction across the different L-S interfaces computed as the sum of hydrogen bonds (using the Mayo potential (Dahiyat *et al.*, 1997)) and hydrophobic interactions (using a hydrophobic potential of mean force (Lin *et al.*, 2007)) for the different stages of the reaction. In Stages I and II, the total interaction energy across the LS1 interface is much stronger than across the two smaller L-S interfaces (LS2 and LS3) while Stages III and IV are characterised by an almost balanced interaction energy for each S subunit with their two adjacent L subunit dimers. This balancing of the energy across the different L-S interfaces (LS1 *vs.* LS2+LS3) in Stages III and IV supports the 'delocalisation' of the S subunits (Figure 6.11) and the 'resonant' structure with enhanced global connectivity. Note that the 'closed' structure (Stage III) has stronger L-S interactions overall. **c.** Location of the subunits and interfaces on the structure.

Hence, changes in the hydrogen bond energies between L and S subunits are responsible for the emergence of an increased global connectivity in the structure at the last two stages of the reaction and upon inhibition through the creation of a communication channel spanning across the quaternary structure via the sharing of small subunits.

6.4 Analysis of large multimers - Application to ATCase & hemoglobin

In this section, we investigate the generality of some of the characteristics of the structural organisation identified in Rubisco and whether they can be transposed to other multimeric structures, especially with observed and well documented cooperative effects. To this end, we conduct a short analysis of two multimeric structures: ATCase, a textbook example of allosteric multimeric enzyme whose catalytic and regulatory mechanisms have been studied extensively, and hemoglobin, a small globular heterotetrameric transport protein and a classic example of cooperativity. In particular, the activity of ATCase is autoregulated by a clever molecular feedback enabled by allostery and homotropic cooperativity.

6.4.1 ATCase

Using a representative dataset of six structures sampling ATCase in different states of allosteric regulation and with different ligands bound, we identify general properties of ATCase's structural organisation and link changes in the intersubunit association patterns to the cooperative mechanisms induced by allosteric effectors.

Structure and function of ATCase

Aspartate transcarbamoylase (ATCase) is the enzyme that realises the first step of the biosynthesis pathway of pyrimidines, one of the two types of nucleic acid bases which includes cytosine, thymine and uracile. ATCase has been extensively studied and has become a classic model of allosteric regulation (see recent reviews by Kantrowitz (2012); Lipscomb and Kantrowitz (2012)).

ATCase is especially known for its ability to regulate its own metabolic pathway by altering its rate of catalysis through both homotropic cooperativity and allostery. Cooperativity is the phenomenon by which the affinity of some binding sites of a

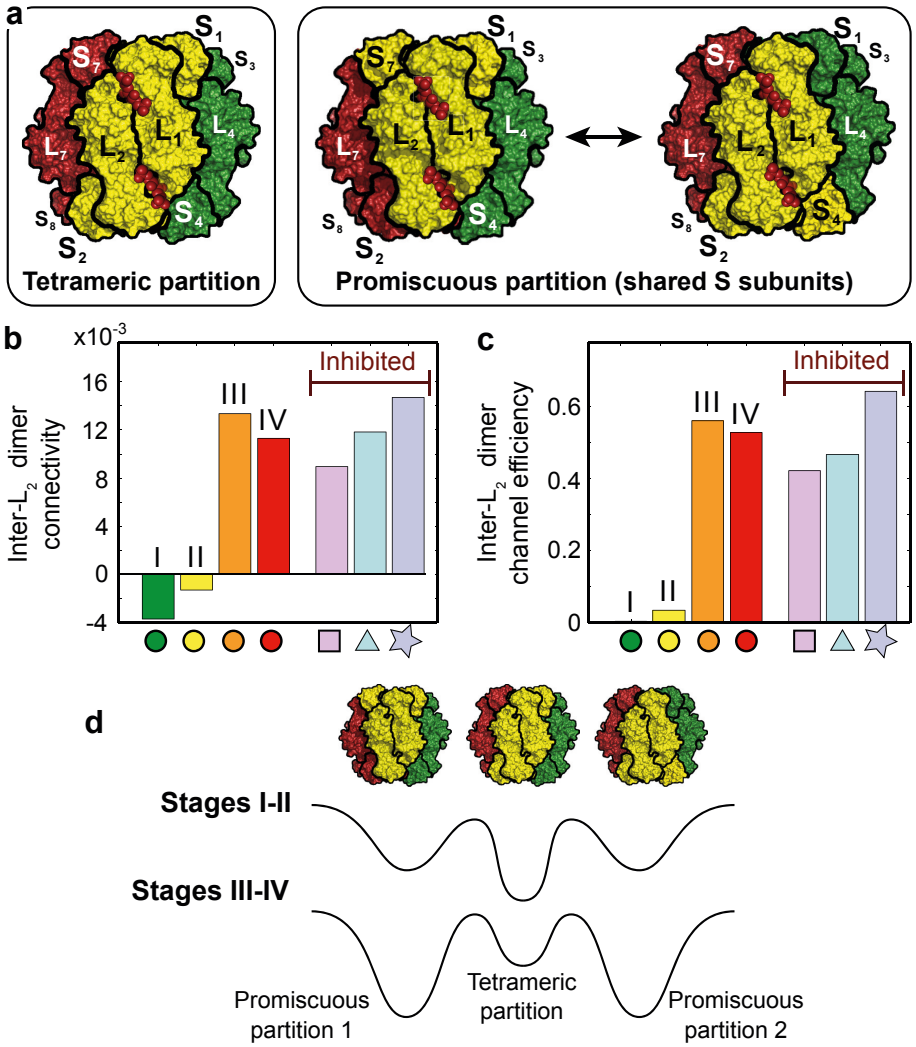


Figure 6.11: Change in the inter L_2-dimer connectivity mediated via the S subunits during the reaction. **a.** The 'tetrameric partition' and the two symmetrical realisations of the 'promiscuous partition', in which the S subunits are shared across dimers. **b.** Inter-L_2 dimer connectivity for all Rubisco structures (same symbols as in Figure 6.9) calculated as the difference in Markov stability between the 'promiscuous' and 'tetrameric' partitions. Stages III and IV (as well as the inhibited structures) exhibit a tendency to share the S subunits across dimers. **c.** Inter-L_2 dimer channel communication efficiency calculated from the probability of the S subunits to be grouped with different L subunits using the information-theoretical channel efficiency defined in Equation (6.2). **d.** Cartoon of the ultimate scale Markov stability landscape at different reaction stages: I and II versus III and IV. The well depths represent the Markov stability computed for each partition shown in **a**.

protein is modified upon binding events in other binding sites of the same protein. It is said to be homotropic if the molecule causing the cooperative effect is also the one being affected by it. In ATCase, homotropic cooperativity is initiated by a major conformational change from a T (tense) state to an R (relaxed) state following the occupation of one active site or so by the substrate (Lipscomb and Kantrowitz, 2012; Macol *et al.*, 2001). The R state favours the binding of additional substrates to the unoccupied binding sites and thus increases enzymatic activity.

ATCase activity is also allosterically controlled by the end products of the purine and pyrimidine pathway. ATCase is inhibited by the binding to the regulatory subunits of cytidine triphosphate (CTP) and uridine-triphosphate (UTP), two end products of the larger pyrimidine pathway it is part of (Kantrowitz, 2012; Cockrell *et al.*, 2013; Cockrell and Kantrowitz, 2012). Similarly, the binding to the regulatory subunit of adenosine triphosphate (ATP), an end product of the purine pathway, stimulates its activity. The binding of these effectors induces a reorientation of key residues within the active sites of the catalytic subunits, and shifts the T/R equilibrium towards respectively the T and R states (Stevens and Lipscomb, 1992; Stevens *et al.*, 1990). These two allosteric effects thus create a feedback control mechanism which allows ATCase to regulate its own metabolic pathway and helps balancing the total amount of purine and pyrimidine nucleotides in the cell.

ATCase catalyses the reaction between apsartate (Asp) and carbamoyl phosphate (CP) to form carbamoyl-aspartate (CA) and inorganic phosphate (Pi). Its structure consists of two trimers of catalytic subunits and three dimers of regulatory subunits (see Figure 6.12). The catalytic subunits comprise two domains associated with the catalytic binding sites for Asp and CP. The regulatory subunit also contains two domains which comprise the allosteric binding sites for the zinc cofactor and the allosteric effector (ATP, CTP or UTP) in the regulatory subunit.

The binding events are ordered, with CP binding before Asp and CA being released before inorganic phosphate (Ke *et al.*, 1988). The binding of CP induces tertiary conformational changes essentially in the 50's, 80's and 240's loops which create a structurally and electrostatically favourable binding site for Asp (Wang *et al.*, 2005). Upon Asp binding, ATCase transitions from a relaxed T state to an excited R state through a major conformational change taking place both locally at the level of the tertiary structure, and globally in the quaternary structure. In particular, this transition involves rearrangements of the 80's and 240's loops which brings the two substrates together and lowers the activation energy of the reaction, and a large quaternary conformational change characterised by an 11 Å elongation

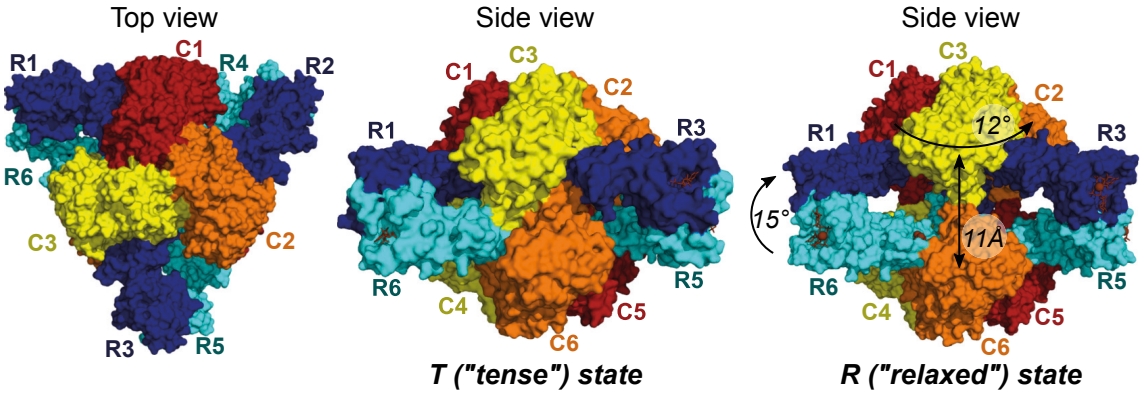


Figure 6.12: ATCase is formed of six catalytic subunits arranged in two trimers linked together by three dimers of regulatory subunits. When the substrate binds one or more of its catalytic sites, ATCase transitions from a T state to an R state. A large conformational change is associated with this transition which includes a 12 ˚ rotation of one catalytic trimer relative to the other, a 15 ˚ rotation of all regulatory dimers, and an 11 Å increase of the distance between the catalytic trimers.

along the 3-fold axis, a 12 ˚ rotation of one catalytic trimer relative to the other triplex, and a 15 ˚ rotation of each regulatory dimer around their 2-fold axis (see Figure 6.12) (Kantrowitz, 2012).

Structural data

We selected a representative subset of six structures of *Escherichia coli* ATCase which capture the conformation of the enzyme at different stages of the reaction and in several states of activation and inhibition by allosteric effectors. In particular, our dataset comprises the unliganded ATCase as well as ATCase bound with a bisubstrate analog, the products (Pi and CA) and each of the allosteric effectors (CTP, ATP and the combination of UTP and CTP in the presence of Mg^{2+}). All structures are bound with the Zn effector in the regulatory subunits. Two of the structures use the bisubstrate analog N-phosphonacetyl-L-aspartate (PALA) which, although effectively inhibiting ATCase by blocking the active sites to which it binds, also mimics the combined structural effect of the two substrates aspartate and carbamyl phosphate (Huang and Lipscomb, 2004) and can thus be considered as the transition state analog (Kantrowitz, 2012).

Intermediate scales and general behaviour

Our analysis of ATCase in Figure 6.13 confirms several key properties of the structural organisation of multimeric structures which we identified in Rubisco.

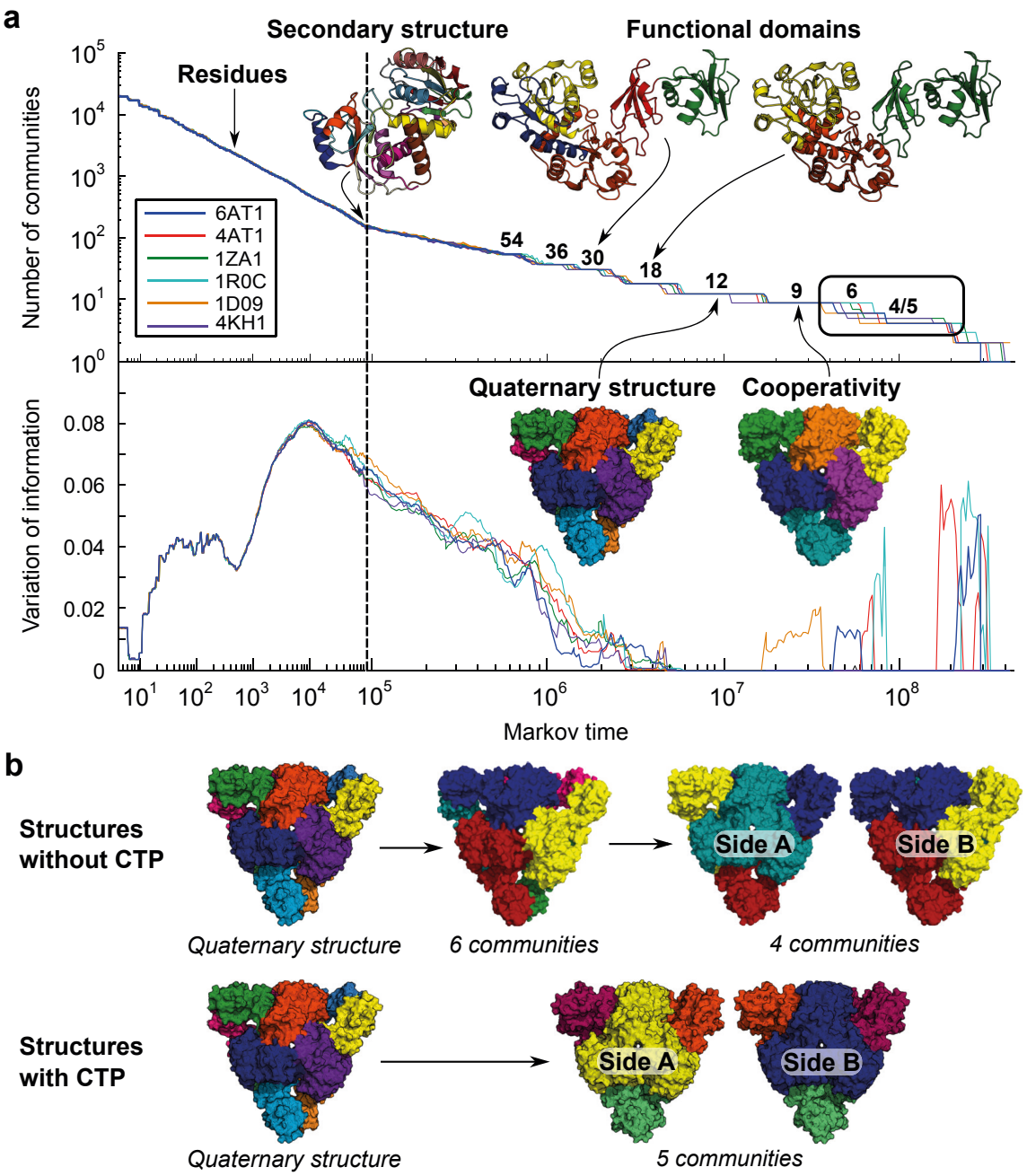


Figure 6.13: Markov stability analysis of ATCase. **a.** Evolution of the number of communities and variation of information with the Markov time for four T state and 2 R state structures of ATCase (see Table 6.1), showing a large number of meaningful partitions in all structures, and a strong similarity across all states and conformations. **b.** Partitions identified in the boxed region of panel **a.**. In the 4, 5 and 6-community partitions, CTP-bound structures distinguish themselves by having strictly hierarchical partitions, while other structures show regulatory subunits being alternatively grouped with another regulatory and a catalytic subunit as Markov time evolves.

PDB	Ligand	State	Reference
6AT1	Zn	T – unliganded	(Stevens *et al.*, 1990)
4AT1	Zn/ATP	T – activated	(Stevens *et al.*, 1990)
1ZA1	Zn/CTP	T – inhibited	(Wang *et al.*, 2005)
1R0C	Zn/Pi/CA	T – product bound	(Huang and Lipscomb, 2004)
1D09	Zn/PAL	R – transition state analog bound	(Jin *et al.*, 1999)
4KH1	Zn/Mg/CTP/ UTP/PAL/Pi	R – inhibited and transition state analog bound	(Cockrell *et al.*, 2013)

Table 6.1: PDB structures of ATCase analysed.

Firstly, similarly to Rubisco and in contrast with the analyses of AdK and MTIP in the preceding two chapters, a larger number of robust levels of organisation are identified in ATCase as indicated by multiple plateaus together with a low variation of information. They reflect the complexity of the functional mechanisms in ATCase as compared to proteins with less elaborate dynamics (see the analysis of hemoglobin in the next section).

Secondly, residues, secondary structure and functional domains are again successfully recovered. In all structures, robust communities obtained within the different subunits perfectly match the asp and CP domains in the catalytic subunits (18 communities at Markov time 2×10^6), as well as the Zn and allosteric domains in the regulatory subunits (30 communities at Markov time 2×10^6).

Thirdly, the community structures are almost undifferentiated between the T and R states despite the large difference in their conformations and the reinforcement and breaking of several key interfaces between subunits (Stevens and Lipscomb, 1992). This result aligns with our previous observations in AdK and supports the model of enzyme dynamics proposed by Henzler-Wildman *et al.* (2007b). The absence of modifications in the global structural organisation of the protein upon binding of substrates or allosteric effectors suggests that the intrinsic organisation of the enzyme at closure is already encoded in the open form. The R and T states appear to differ only in the 18-community partition which is marked by a slightly smaller plateau in the R state (structures 1D09 and 4KH1). This partition corresponds to the detection of the two domains of the catalytic subunits (12 communities), and the regulatory subunits (6 communities). The smaller robustness of this partition

thus indicates a smaller disconnectivity between the binding domains of the two substrates in the catalytic subunits, which could reflect a slightly enhanced communication between the pairs of catalytic subunits, in parallel with the stronger association we observed between the two lobes of Rubisco active site in the closed conformation.

Large scale organisation and impact on cooperativity

Beyond these general properties of proteins and multimeric structures, our analysis also unveils possible links between the structural organisation of ATCase and the mechanisms of allosteric and homotropic cooperativity. At Markov time 6×10^6, the partition into 12 communities successfully identifies the quaternary structure of ATCase, each community englobing a single subunit. At longer Markov times, favoured subunit-subunit associations are revealed in the subsequent coarser partitions which signal particular patterns of communication across the whole multimer.

The first intersubunit partition to appear (9 communities) groups together the pairs of regulatory subunits into the three dimers. The identification of the regulatory dimers before the catalytic trimers and the dimers of catalytic and regulatory subunits is indicative of a particularly strong communication between them. This partition is also the first to group subunits from the two sides of the multimer, and thereby underlines the key role played by the regulatory subunits in transmitting the necessary signals for homotropic cooperativiy between the two catalytic trimers (Stevens and Lipscomb, 1992; Newell *et al.*, 1989). This partition is identical in all structure except for ATCase bound with the transition state analog PALA. In this structure, the three regulatory subunits from one side only are dissociated: their Zn domain joins the neighbouring catalytic subunit while the allosteric domain is grouped into one community with the other regulatory subunit. Although the reasons for this disconnectivity are unclear, previous molecular dynamics simulations conducted on the same state of PALA-bound ATCase (Tanner *et al.*, 1993) found the allosteric domains to be mechanically uncoupled from the zinc domains in this state, which agrees with our results.

The next partition into six communities groups together the catalytic subunits with their associated regulatory subunit. Although this partition seems natural, it is surprisingly short-lived and non robust, as indicated by the variation of information and the small plateau, and is even absent in the structures of ATCase inhibited by CTP. Regulatory subunits are thus weakly linked to their catalytic subunit comparatively to their paired regulatory subunit.

The partitions into four and five communities highlight changes in the intersubunit communication patterns upon allosteric inhibition. Except for the CTP-bound ATCase, all structures follow the subunit grouping pattern in the upper panel of Figure 6.13b. Following the partition into six communities, subunits are associated following an asymetric four community partition containing three communities formed by the regulatory dimers with one catalytic subunit from one side of the structure, and one community containing the catalytic trimer from the other side of the structure. Unlike Rubisco, the symmetric equivalent of this partition has a much lower Markov stability and never appears as a suboptimal solution in the Louvain ensemble of solutions.

This result, together with the six community partition found in the 9 community partition of PALA-bound ATCase, actually reflects the inherent asymmetry of ATCase. Previous crystallographic studies identified structural asymmetries in the regulatory subunits in PALA-bound ATCase (Jin *et al.*, 1999) and CTP-bound structures (Kim *et al.*, 1987; Lipscomb and Kantrowitz, 2012; Kosman *et al.*, 1993), which are thought to be preexisting in other conformations as well (Kantrowitz, 2012), as well as differences in the nucleotide binding constants of the two regulatory chains (Winlund and J. Chamberlin, 1970; Mendes *et al.*, 2010).

Our results suggest that these assymmetries impact ATCase differently in different states. Indeed, we found that structures inhibited by CTP exhibited a different pattern of association between subunits at these scales, as shown in the lower panel of Figure 6.13b. Unlike the other structures, the partitions are here perfectly symmetric, and transition directly from the identification of the regulatory dimers to the partitioning into the three regulatory dimers and the two catalytic trimers.

An important difference between the two sequences shown in Figure 6.13 is the sharing of the regulatory subunits through scales. In the upper panel, the regulatory subunits are alternatively grouped with the catalytic subunit and into dimers while in the CTP-bound structures, the groupings follow a strict hierarchy. This once again relates to the results obtained for Rubisco at the largest scales. Although we do not observe a sharing of subunits at the same level of organisation within the Louvain ensemble, the sharing takes place here at different scales. Rather than both types of partitions being both equally relevant at a same Markov time, they are here alternatively optimal at different scales. Similarly to Rubisco, this sharing could create a communication channel between the two catalytic trimers which could enable the homotropic cooperativity between them.

Unlike the other structures, CTP-bound ATCase appears to have no sharing of the regulatory subunits in our partitions. This suggests a decreased level of communication between the catalytic trimers in the presence of CTP which may negatively impact homotropic cooperativity. Interestingly, this reduced communication level could explain the decrease in cooperativity induced by CTP observed by Newell *et al.* (1989) in binding assays of CTP-bound ATCase with PALA.

In conclusion, our results suggest that the small local asymmetries in the structures of ATCase propagate upwards at quaternary level where they have a direct global impact on the communication between subunits which could influence cooperative and allosteric mechanisms.

6.4.2 Unstructured biomolecules - Application to hemoglobin

The rich multiscale hierarchical organisation of Rubisco and ATCase identified using Markov stability is in sharp contrast to that rendered by other proteins with simpler structure and dynamics. In this section, we oppose the intricate and refined structural organisation of Rubisco to hemoglobin, a well understood globular protein with an unsophisticated tertiary structure and simple dynamical behaviour.

Structure and function of hemoglobin

An essential protein for almost all vertebrates, hemoglobin is a small globular multimer which transports oxygen and carbon dioxyde between the lungs and organs in the bloodstream. As the protein where cooperativity was first discovered, it is one of the most extensively studied protein and the first to have its three-dimensional structure resolved.

In human adults, its tetrameric structure contains two types of subunits, α and β, arranged in two identical heterodimers. α and β subunits are very similar in sequence and structure and consist of a single domain all-α-helical globin fold comprising respectively seven and eight α-helices. Each subunit encloses one heme group, an aromatic chemical compound containing an iron ion which binds oxygen.

Hemoglogin is the paradigm for homotropic cooperativity in proteins and has been the subject of a long history of models (reviewed by Eaton *et al.* (1999, 2007)). Perutz (1970) was the first to understand the structural basis for cooperativity. Similarly to ATCase, hemoglobin exits in two states: the deoxy T state (no oxygen), and the oxy R state (oxygen-bound) characterised by a higher affinity for oxygen. Oxygen binding shifts the equilibrium towards the R state, and induces

local rearrangements in the individual subunits. These local changes in the tertiarty structure propagate to the other subunits by altering the α_1/β_2 interface, which triggers a global conformational change in the quaternary structure that can be approximated by a 15 $^\circ$ rigid body rotation and 1Å translation along the rotation axis of one α/β dimer with respect to the other.

In the context of this work, hemoglobin serves as an ideal counterexample to the rich multiscale structural organisation we identified in ATCase and Rubisco. Hemoglobin is a single domain globular protein without any clear structural organisation beyond the secondary structure, and the T-R conformational change mostly takes place in the quaternary structure with only minor tertiary rearrangements. The structural organisation of hemoglobin outside of the secondary and quaternary structures should thus be minimal.

Markov stability analysis of hemoglobin

Figure 6.14 shows a comparison of the results for activated unliganded spinach Rubisco and human T state hemoglobin (PDB 1GZX). At small scales (from Markov time 1 to 10^6), both proteins show the same organisation since Markov stability recovers the basic chemistry and biochemistry, including the chemical groups, residues and secondary structure, which is shared by all proteins. At intermediate scales and beyond, however, both proteins show strikingly different organisations. In the case of hemoglobin, only the four subunits that compose the quaternary structure are identified as a relevant community structure—with the exception of a partition into eight communities showing a small plateau but associated with a much higher variation of information. Rubisco, on the other hand, shows the complex hierarchy described in the previous section with nine plateaus in the number of communities, eight of which are associated with a low variation of information and are linked to key aspects of Rubisco's functional mechanisms.

We have also tested the relevance of the structural organisation at large scales for Rubisco and hemoglobin through another significance test. This is achieved through a comparison against an ensemble of surrogate random graphs using the weak interactions model described in Chapter 4. The randomisation is here aimed at the secondary and tertiary structure and perturbs all the weak interactions (hydrogen bonds, hydrophobic tethers and salt bridges) present in the original structure between atoms separated by more than four residues along the sequence. In order to preserve the quaternary structure, intra-unit and inter-unit weak interactions are randomised independently. This scheme thus effectively randomises the structural

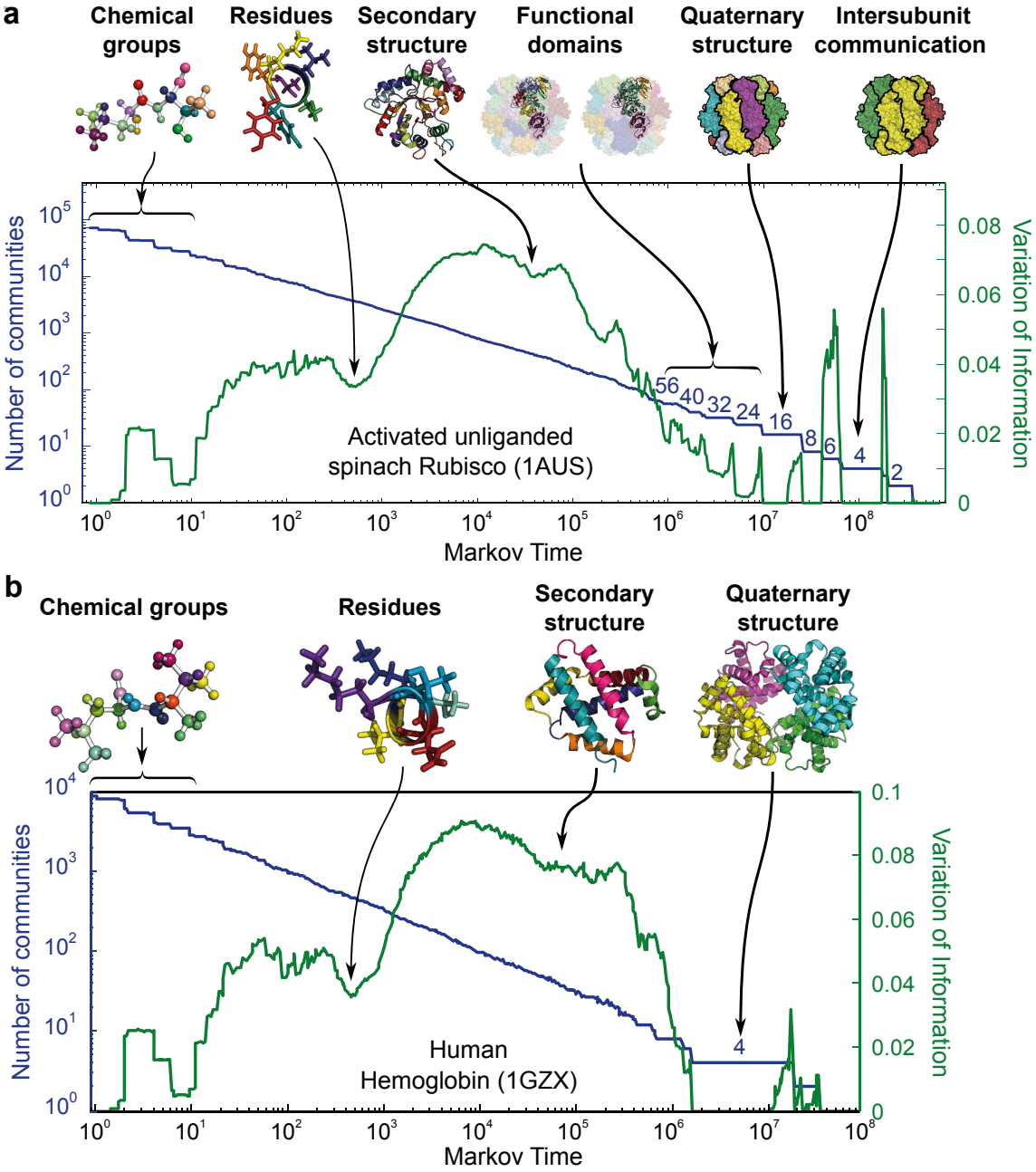


Figure 6.14: Comparison between the all-scale Markov stability analysis of activated unliganded spinach Rubisco (1AUS, in **a**) and T state human hemoglobin (1GZX, in **b**). The chemical and biochemical constituents identified at small Markov times (up to 10^6) are similar in both cases and lead to a very similar Markov time evolution up to those scales. However, the two proteins show distinct outcomes at larger Markov times. The limited number of local minima in VI (green) and long plateaus in the number of communities (blue) suggests that hemoglobin has a simpler multiscale organisation, with only individual subunits identified as a meaningful level of organisation (long-lived 4-community partition).

organisation of the protein beyond helical turns within each subunit, and also blurs the details of the inter-subunit interactions while keeping the total interaction energy constant between all pairs of subunits. Figure 6.15 shows that the hierarchical organisation of Rubisco at large scales exhibits a marked difference against the surrogate models, while hemoglobin remains very close to the ensemble of randomised models. The structure of hemoglobin is thus well approximated by four chains of randomly interacting amino acids which indicates a very low degree of organisation at the tertiary level.

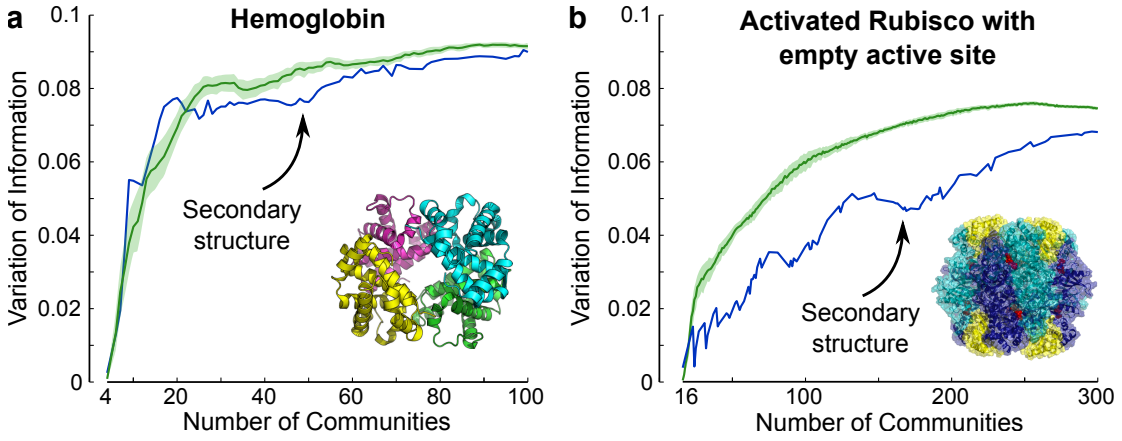


Figure 6.15: Variation of information (VI) of the partitions found for T state human hemoglobin (1GZX, in **a**) and activated unliganded spinach Rubisco (1AUS, in **b**). The VI of the protein structures (in blue) is compared to the VI of an ensemble of random graph surrogates (in green) in which the higher scale structural organisation has been blurred by relocating at random the intra-subunit weak interactions beyond the secondary structure (see text). Green lines and shaded area correspond to the mean and standard deviation of an ensemble of 100 surrogate random graphs. Rubisco exhibits a clear difference between the protein and its surrogates, indicative of a much richer and better defined structural organisation as compared to a globular protein such as hemoglobin, in which the randomisation beyond the secondary structure is indistinguishable from the original protein.

6.5 Discussion

The low computational cost of Markov stability not only permits the analysis of very large structures such as ATCase and Rubisco, but also the identification of structural features linked to dynamical events over the longest time and spatial scales, including collective behaviours involving multiple subunits and spanning entire multimers. In this chapter, we have used Markov stability to uncover bottom-up, the structural anatomy of Rubisco at all scales starting from a fully atomic descrip-

tion, and avoiding re-parameterisation or *a priori* coarse-graining often needed in such large biomolecules. As an important and well-researched enzyme, Rubisco provides a suitable stage to demonstrate the power of the method and to enlighten our knowledge of the workings of complex multimeric structures. The generality of our approach and conclusions was also shown through a comparison with ATCase, a classic example of a multimeric enzyme with cooperative and allosteric mechanisms and of similar complexity to Rubisco, and hemoglobin, a globular multimeric transport protein with a very simple structural organisation.

Our analysis of the structural data of spinach Rubisco throughout the catalytic reaction (Stage I: "activated unliganded analog"; Stage II: "activated substrate bound"; Stage III: "transition state analog"; and Stage IV: "activated product bound") as well as with three different inhibitors bound sensitively unveils changes throughout the reaction involving not only substrate-induced intra-unit domain movements but also communication patterns over the whole structure at the level of the quaternary structure. At intermediate scales, we find that the α/β barrel is organised in two possible hierarchies (Types I and II in Figure 6.8) and that both hierarchies are encoded as fingerprints in the structure. However, the two hierarchies appear with different likelihood throughout the steps of the catalytic reaction, suggesting a switch from one to the other in the course of the reaction. In particular, both the 'open' activated/unbound conformation (Stage I) and the 'closed' transition state conformation (Stage III) have the same hierarchical organisation (Type I) but with increased connectivity between key structural areas (loop 6 and anchoring points) in Stage III. The intermediate steps (Stages II and IV) involve rearrangements that lead to a more fluid hierarchical organisation (Type II) with enhanced integrity in the vicinity of each structural element linked to the conformational change (phosphate anchoring subsites P1/P2 and loop 6) (Duff *et al.*, 2000). We find that the rearrangements of the subunits also have effects beyond the single unit. We find that the change in the interactions of the α/β barrel with the N-terminal domain of the *neighbouring* subunit (Duff *et al.*, 2000) associated with closure of the structure leads to increased inter-unit connectivity between the two lobes of the active site within the L_2 dimer (Figure 6.9) in the "closed" conformation.

Our analysis also reveals a role for the S subunits in enhancing the communication across all the L_2 dimers in the structure. Despite extensive research into Rubisco, the function of the S subunits remains enigmatic. It is well known that the L_2 dimer is the minimal catalytic processing unit and the S subunit is not essential for catalysis (Tabita *et al.*, 2008; Gutteridge, 1991; Hartman and Harpel, 1994).

Other multimeric enzymes of comparable complexity, such as ATCase, have been shown to make use of non-catalytic subunits, which are often smaller, as regulatory domains to control their efficiency. Regulatory inputs are most often triggered by an allosteric effector binding to the regulatory subunit, causing binding sites to influence each other and resulting in the observed cooperativity. Indeed, the S subunits have been hypothetically associated with cooperativity in Rubisco (Knight *et al.*, 1990; Yokota *et al.*, 1991; Taylor and Andersson, 1996). Remarkably, our method finds that Stages III and IV, as well as inhibited structures, are characterised by a change in the relative strength of hydrogen bonds between the L-S subunits leading to increased symmetry of the interactions. These subtle changes lead to the "resonant" communities in which the S subunits are shared across the L_2 dimers with increased connectivity across all the L_2 dimers at longer time scales mediated via the S subunits (Figure 6.11). This increased connectivity could be indicative of a direct involvement of the S subunits in the completion of the reaction at time scales that are much slower than the loop 6 rearrangements associated with the 'closing' of the structure. We observed a similar effect in ATCase, whose cooperative mechanisms are very well known. In our analysis, regulatory subunits were indeed alternatively associated with catalytic subunits from different trimers as the Markov time evolved. This effect was notably absent in CTP-bound structures which are known to have a reduced cooperativity between the trimers.

Intriguingly, the multiscale organisation of Rubisco is remarkably well-defined and robust; much more so than most proteins, as showed our analysis of hemoglobin, which displays a less pronounced hierarchical organisation. This suggests that the complexity of the functional mechanisms of a protein are reflected in its multiscale structural organisation. Conversely, ATCase, whose multistep reaction and diverse cooperative mechanisms are well known, displayed a complex hierarchy of scales similar to Rubisco.

Long range effects such as allostery and cooperativity, whereby the binding of a molecule can propagate and impact events at very distant catalytic sites, remain a central but poorly understood phenomenon in biochemistry. In multimeric structures, cooperative and allosteric signals often spread from one subunit to another, but identifying the channels of communication remains a difficult problem. Although our method is not aimed at revealing the path taken by allosteric signals throughout the structure, our results identified preferred intersubunit associations which shed light on the way subunits interact and possibly communicate in multimeric structures.

Chapter 7

DNA quadruplexes

So far, only proteins have been considered. Due to its general applicability, this methodology however extends to other biomolecules with a distinct biological role and dynamical behaviour. Nucleic acid structures, in contrast to proteins, are governed by a different mix of physico-chemical interactions, with electrostatic and π-stacking interactions dominating along with hydrogen bonds. These, in turn, yield unique dynamical as well as structural properties which have crucial implications for their function.

Whereas DNA has long been seen as little more than an information storage medium for the cell, its structural and mechanical properties have now been found to be essential to many physiological processes. DNA indeed shows a certain degree of conformational inhomogeneity that departs from the double helix model of Watson and Crick. Crucially, the structural details of its energetically favourable conformations determine key aspects of its interactions with proteins such as transcription factors and histones which govern some of the most essential biological processes including DNA packaging, repair and replication, as well as gene expression. Beyond the biological significance of its *in vivo* structure and dynamics, DNA is also becoming a prominent material in nanotechnology, and is now increasingly used as a building block for the design of new forms of nanostructures.

Beyond the classical Watson and Crick double helix, DNA thus adopts a wide range of other topologies *in vivo*. In this chapter, we use Markov stability to analyse G-quadruplexes, characterised by the association of four strands of DNA driven by the formation of planar quartets of guanine bases. DNA quadruplexes possess a highly polymorphic nature, and interconvert between multiple conformations dynamically. The challenge is here to generalise our approach to other types of

biomolecules and establish whether signatures of a structural organisation can also be identified in biomolecules with very different structural properties and dynamics such as DNA, and linked to biologically relevant mechanisms.

7.1 Structure and function of DNA quadruplexes

For more than 50 years, guanine rich DNA sequences have been known to form planar tetrads of guanine bases linked together by Hoogsten base pairing (Gellert *et al.*, 1962). Later, the discovery that they arrange DNA into quadruple stranded structures and could form in biologically important region of the genome such as promoter[1], and telomeric regions[2] has sparked a new surge of interest in their biophysical properties (Moyzis *et al.*, 1988; Sen and Gilbert, 1988; Williamson *et al.*, 1989). G-quadruplexes are now thought to be involved in a variety of biological processes (Maizels, 2006; Bochman *et al.*, 2012). Cell death and cancer in particular have been found to be related to their formation at the telomeric ends of the chromosomes and quadruplexes are now increasingly considered as preferential targets for anti-cancer drugs (Stewart and Weinberg, 2006; Neidle, 2010). In addition to their biological importance, their mechanical, self-assembly and electron transport properties make them ideal building blocks for DNA nanotechnology (Alberti *et al.*, 2006; Tran *et al.*, 2013)

G-quadruplexes can be broadly defined as four-stranded nucleic acid structures held together by a core of at least two stacked tetrads of guanine bases that are stabilised by monovalent ions (Neidle and Balasubramanian, 2006; Neidle, 2009). They can result from the folding and association of a single, two or four separate strands of DNA. When formed by a single or two individual DNA molecules, the different strands are linked by loops which vary in length and conformations (see Figure 7.1). Depending on the sequence, cation type and concentration, bi- and unimolecular quadruplexes can consequently display a wide range of different topologies

[1]Region of the DNA where a gene transcription is initiated.

[2]Telomeres are repeated sequence motifs at the end of the chromosomes which protect them from erosion or fusion events with other chromosomes. At every cell division, when the DNA is duplicated, the telomeric ends get shortened. They are however maintained by the enzyme telomerase which adds multiple copies of the telomeric sequence motif. With every cell division event, the total length of the telomeric region nonetheless decreases and thereby limits the number of cell divisions. In cancer cells, several mechanisms are involved which maintain the length of the telomeric ends such as the overexpression of the telomerase enzyme. The stabilisation of the quadruplexes at the telomeric ends has been found to inhibit telomerase activity, and quadruplexes are consequently seen as possible targets for antitumor agents (De Cian *et al.*, 2008; Shay and Wright, 2011).

distinguished by the individual orientation of each strand, the configuration of their connecting loops and the orientation of the individual guanine bases (Phan *et al.*, 2006; Burge *et al.*, 2006). In addition, quadruplexes formed by the human telomeric sequence 5'-AGGG(TTAGGG)$_N$-3' (and variants) have been found to be highly polymorphic, with at least six different topologies observed experimentally to date (Lee *et al.*, 2005; Phan, 2010) between which telomeric sequences are thought to interconvert dynamically in solution (Xue *et al.*, 2011; Dai *et al.*, 2008). Finally, quadruplex folding is kinetically complex and is thought to proceed through a multistep process with several stable intermediates (Lane *et al.*, 2008).

Although the equilibrium conformation has been well characterised for a large number of G-quadruplexes by X-ray and NMR structural data, their dynamical behaviour is still largely unknown. Yet their ability to easily change conformation is likely to be linked to their biological function and understanding their dynamics is therefore critical for the design of more effective drugs (Xue *et al.*, 2011; Zhang and Balasubramanian, 2012). Many biological processes are indeed regulated by kinetic control and operate on transient rather than fully equilibrated conformations. Drugs designed solely on the fully folded structure may thus fail to target these biologically relevant intermediates.

A number of recent studies points towards quadruplex folding/unfolding proceeding via a multistep pathway with kinetically significant intermediates, which could include hairpins, triplexes or different transient quadruplex folds (Chaires, 2010; Lane *et al.*, 2008; Zhang *et al.*, 2010; Stadlbauer *et al.*, 2013; Gray and Chaires, 2008; Mashimo *et al.*, 2010; Bončina *et al.*, 2012; Li *et al.*, 2013; Gray *et al.*, 2014). However, the structural details of the relevant intermediates along the pathways through which quadruplexes fold, unfold and interconvert between topologies are not yet clearly understood. Indeed, the presence of multiple conformations in solution and slow folding kinetics make experimental analysis difficult, while traditional computer simulation techniques remain too limited in the time scales they can access to fully resolve folding and unfolding events and explore quadruplexes polymorphism in depth.

In this work, we aim at bridging this gap by relating the structure of quadruplexes at equilibrium to their unfolding process. To this end, we make use of Markov stability to decompose the global structure of quadruplexes into a hierarchy of clusters of tightly interconnected atoms that are likely to display a collective behaviour during the conformational transitions. Through the analysis of the all-scale structural organisation of an ensemble of nineteen different quadruplex structures, we

a

	PDB ID	Method	Ion	Sequence	Topology	Tetrads
Human telomeres / *Unimolecular*	143D	NMR	Na+	$A(G_3T_2A)_3G_3$	Basket	3
	1KF1	X-Ray	K+	$A(G_3T_2A)_3G_3$	Propeller	3
	2LD8	NMR	K+	$(G_3T_2A)_3G_3T_2$	Propeller	3
	2GKU	NMR	K+	$T_2(G_3T_2A)_3G_3A$	Form 1	3
	2JSM	NMR	K+	$TA(G_3T_2A)_3G_3$	Form 1	3
	2JSL	NMR	K+	$TA(G_3T_2A)_3G_3T_2$	Form 2	3
	2JPZ	NMR	K+	$T_2A(G_3T_2A)_3G_3T_2$	Form 2	3
	2KKA	NMR	K+	$A(G_3T_2A)_3G_3T^*$	Form 3	2
	2KF8	NMR	K+	$(G_3T_2A)_3G_3T$	Form 3	2
	2KM3	NMR	K+	$A(G_3CTA)_3G_3$	Chair	2
	186D	NMR	Na+	$(T_2G_4)_4$	"Form 2"	3
	1I34	NMR	Na+	$G_2T_4G_2CAG_3T_4G_2$		2
Bimolecular	156D	NMR	Na+	$G_4T_4G_4$	Diagonal	4
	1JPQ	X-Ray	K+	$G_4BrT_3G_4$	Diagonal	4
	1K4X	NMR	K+	$G_4T_4G_4$	Diagonal	4
	1K8P	X-Ray	K+	$BrAG_3BrTAG_3T$		3
Tetra-molecular	1NP9	NMR	K+	T_2AG_3		3
	352D	X-Ray	Na+	TG_4T		4
	3TVB	X-Ray	Na+	G_4		4

b

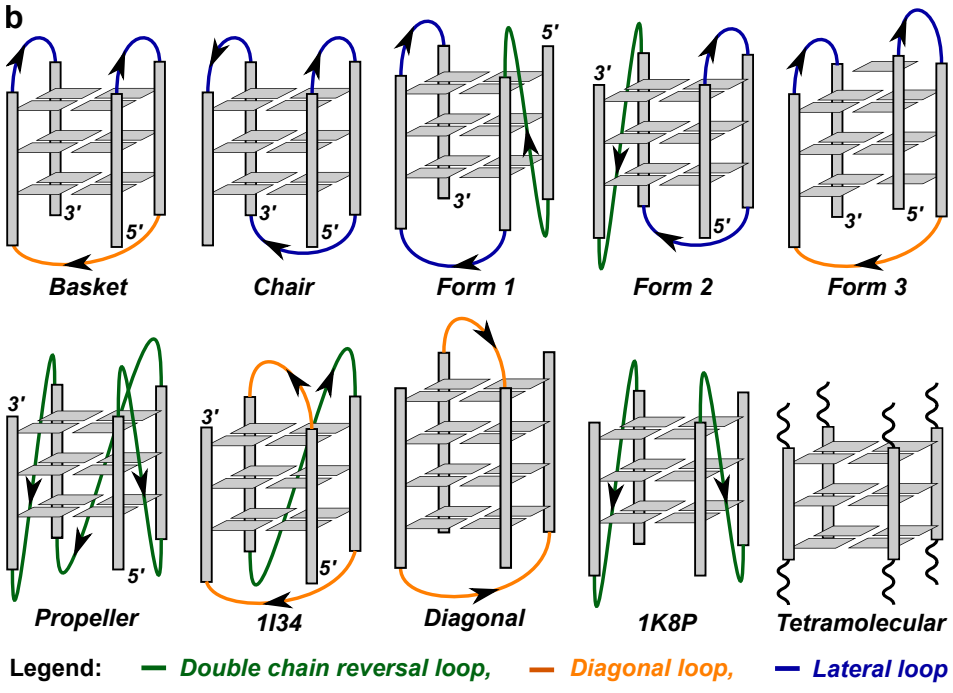


Figure 7.1: **a.** Details of the dataset of nineteen PDB structures used in this work. **b.** The ten different topologies included in the dataset. Three different types of loops can be found in the structures: lateral (in blue), diagonal (in orange) and double chain reversal (in green).

hope to get a better understanding of their general dynamical properties, identify the structural characteristics that enable their high polymorphism, and determine likely pathways of conformational transitions between the different topologies.

7.2 Results

We here analyse and compare the multiscale structural organisation of nineteen single, double and quadruple stranded DNA quadruplexes, nine of which are human telomeric DNA structures, and which encompass ten different topologies (Figure 7.1). We show that our methodology recovers the elementary chemical and biochemical organisation of DNA, and identifies the bases involved in the guanine tetrads core of the quadruplex. It also captures the importance of ions in maintaining the physical stability of the quadruplexes, and provides indications on the origin of quadruplexes polymorphism and their physical stability. Finally, we show that signatures of the unfolding process are already encoded within (folded) structures: the first bases to unfold show a more pronounced disconnectivity from the quadruplex G-tetrads core and we exemplify this property of our method in the case of the human telomeric DNA interconversion between different quadruplex topologies.

7.2.1 Structural data

The dataset of G-quadruplex structures used in this study is summarised in Figure 7.1. When multiple structural models were included in the PDB file, as is often the case for NMR data, each model has been analysed separately and the partitions presented here are taken from the most representative example of the dataset. When ions were missing from the PDB file[3], they have been manually included in the structure: Na^+ ions are positioned in plane with the G-tetrads, while K^+ ions are placed in-between pairs of tetrads, in agreement with their position in X-ray structures of quadruplexes (Neidle and Balasubramanian, 2006).

7.2.2 The all-scale structural organisation of DNA quadruplexes

In Figure 7.2a, we illustrate our approach with the all-scale structural analysis of the propeller type human telomeric quadruplex (1KF1). As we have previously observed in proteins, Markov stability identifies the elementary chemical and biochemical building blocks at small Markov times. In DNA structures, the chemical

[3]Ions often cannot be resolved by NMR spectroscopy.

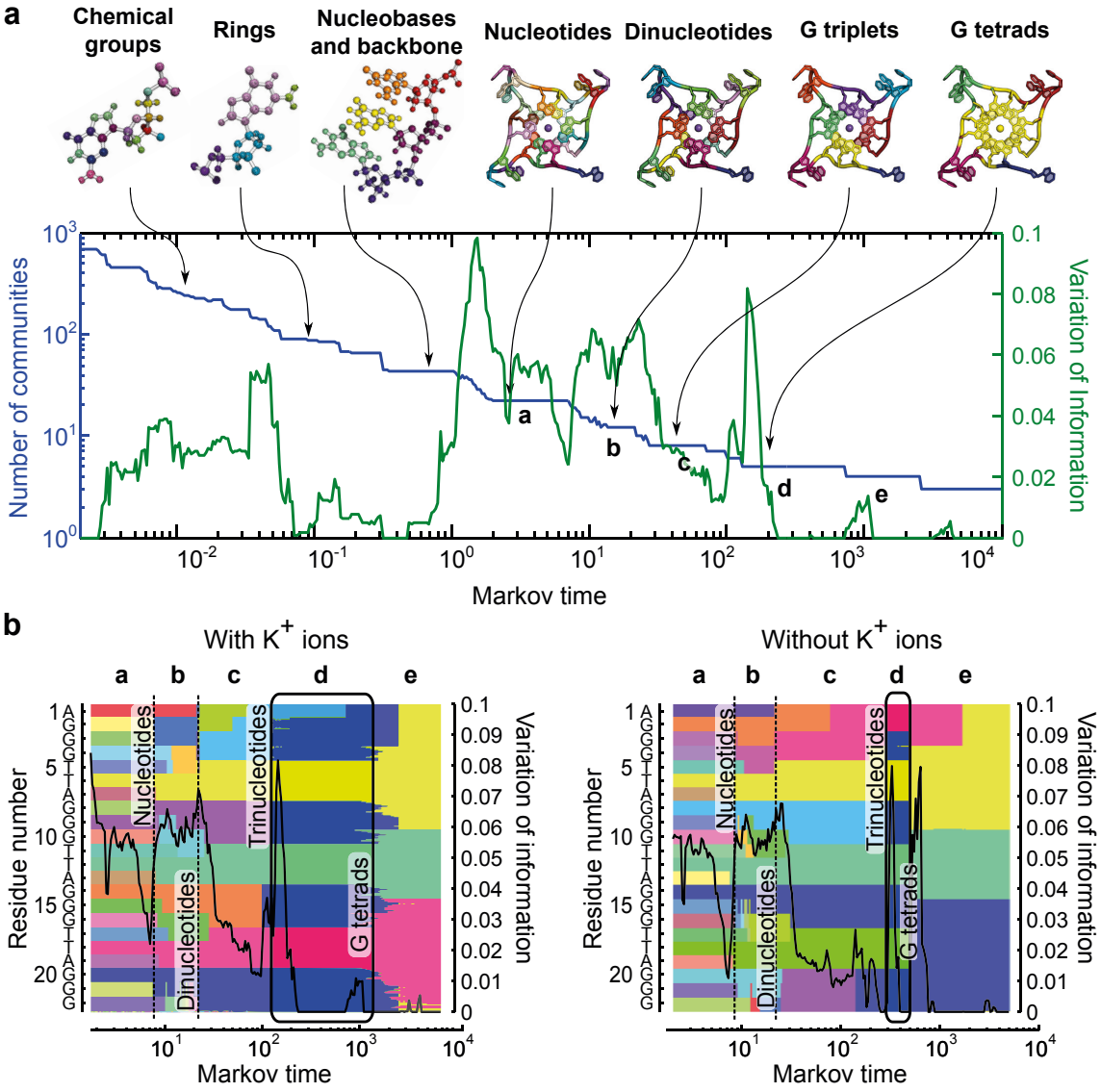


Figure 7.2: **a.** Markov stability analysis of the d[A(G_3T_2A)$_3G_3$] unimolecular propeller quadruplex (PDB 1KF1). The basic chemical (chemical groups and aromatic rings) and biochemical building blocks (nucleobases and nucleotides) are identified at small scales, followed by groupings of two and three nucleotides and a community englobing the three G-tetrads. All are associated with a strong persistence over Markov time (long plateaus in the number of communities) as well as an increased robustness (sharp decrease in the variation of information). **b.** The two charts show the Markov time evolution of the partitioning of the propeller quadruplex with and without the inclusion of the stabilising potassium ions. The left y axis lists the different residues in order and the colours indicate the community membership of each residue. The comparison between the two charts shows that the removal of the stabilising ions significantly disrupts the structural organisation of the quadruplexes, in particular at the level of the G-tetrads community whose Markov time persistence is considerably decreased. In the other unimolecular structures analysed, the tetrads community actually completely disappears in the absence of ions.

groups, aromatic rings, nucleobases, backbone (formed by the deoxyribose and phosphate groups), and nucleotides are all identified as significant community structures (presence of a plateau in the number of communities and drop in the variation of information). As the Markov time grows, larger groups of two and three nucleotides are identified, leading to the formation of a single community by the three G-tetrads.

These levels of organisation are common to all the DNA quadruplex structures analysed here (see Figure 7.3 where the results for all unimolecular structures are superimposed). As expected, the different structures are virtually indistinguishable by our analysis up to the individual nucelotides since all share the same elementary chemical and biochemical building blocks. Similarly, the identification of all the G-tetrads as a single group, well separated from the loops, is common to all the quadruplex structures analysed. Yet the path through which individual nucleotides aggregate to form the G-tetrads and beyond differ amongst structures in terms of the scale at which groupings appear, their robustness, and the order in which individual nucleotides associate with each other.

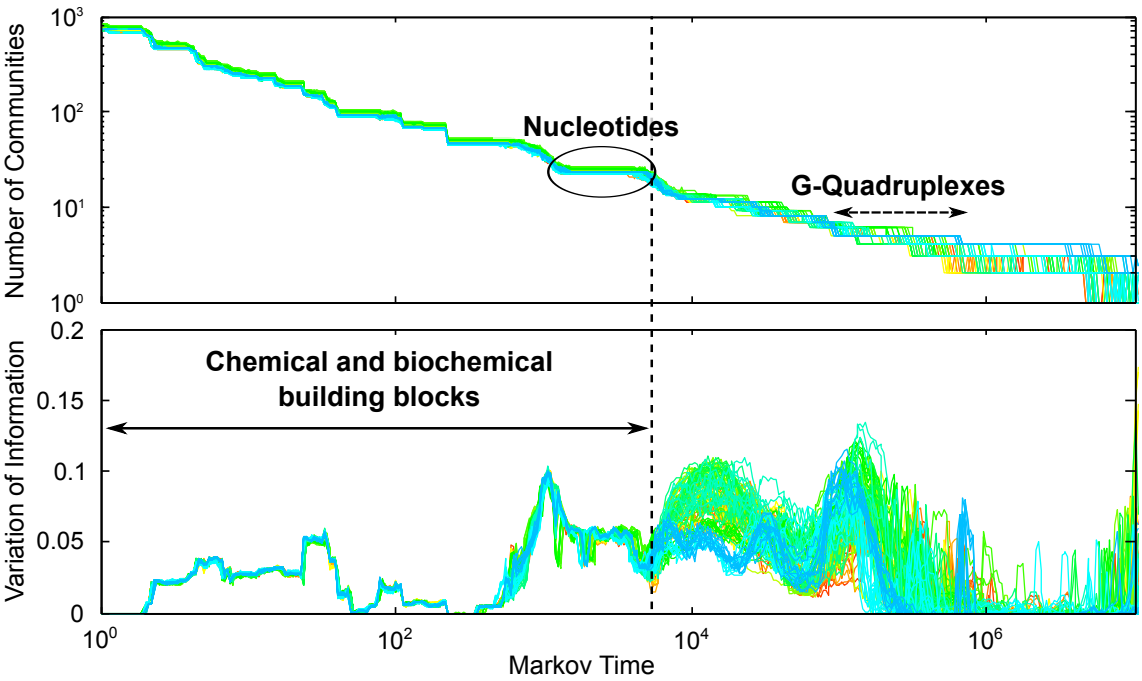


Figure 7.3: Markov stability analysis of all unimolecular DNA quadruplexes. All quadruplexes remain undifferentiated in their structural organisation at early Markov times as they are made up of the same chemical and biochemical building blocks. From the intermediate scales onwards, they display a great variability, yet, at larger scales, all quadruplexes show with high robustness a community formed by the ensemble of G-tetrads.

7.2.3 Role of the stabilising cations

Monovalent cations such as Na^+ and K^+ usually sit either in plane or in between the tetrads and are coordinated by the carbonyl oxygens of the guanine bases. Their presence is a key requirement for quadruplex formation and efficient energy transfer (Dumas and Luedtke, 2010), and their type and concentration has a considerable impact on the physical stability of the quadruplexes (Hud and Plavec, 2006). To estimate their influence on the structural organisation of the quadruplexes, we reconstructed the graph of each quadruplex after having removed the coordinating cations from their original structure.

Figure 7.2b shows a comparison of the Markov stability analysis of the propeller type human telomeric quadruplex (1KF1) with and without the presence of the central cations, from the identifications of the nucleotides onwards. Although most of the communities remain the same in both cases, the lifetime of the G-tetrads community is considerably reduced. A similar effect is observed for all the structures analysed, the tetramolecular structures being the least sensitive and the unimolecular quadruplexes the most sensitive to the removal of the cations. Except for the propeller form, the absence of stabilising cations even resulted in the complete extinction of the tetrads community in all other unimolecular structures analysed. In agreement with experimental observations, we thus find the presence of the cations to have a crucial impact on the formation of a robust community by the core of guanine tetrads which stabilises the quadruplex structure. In addition, we find the influence of the cations to be limited to the structural organisation of the quadruplexes at the largest scales, i.e. from the identification of the tetrads community onwards.

7.2.4 Bases involved in the G-tetrads

Although the majority of quadruplex structures involve the maximum number of guanine bases in the formation of tetrads, structures exist where the number of tetrads does not actually reflect the number of guanine bases present in the sequence. This is notably the case for the unimolecular quadruplexes form 3, chair type and $(T_2G_4)_4$ (2KKA, 2KF8, 2KM3 and 186D, see Table 7.1). In Figure 7.4, we show that Markov stability correctly identifies three tetrads in $(T_2G_4)_4$, and only two tetrads in human telomeric form 3. For each structure, only the bases actually involved in the tetrads are included in the G-tetrads community by Markov stability, irrespective of their type or the stacking interactions they form, and our methodology is thus

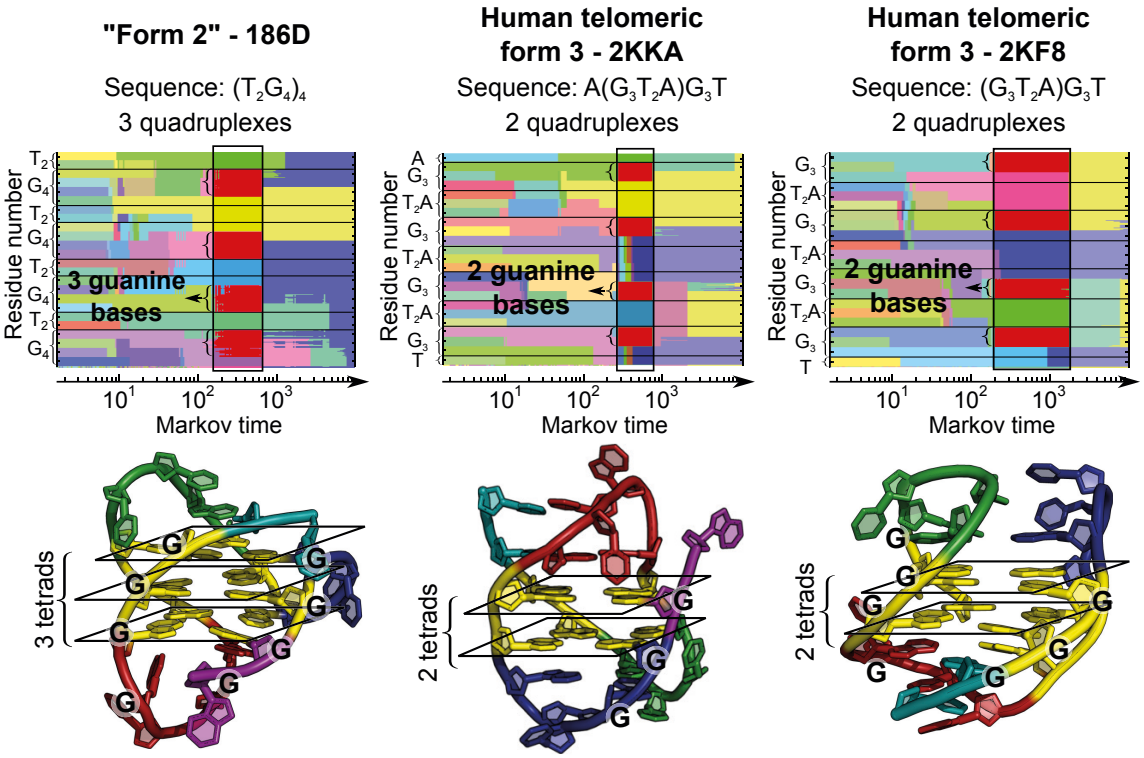


Figure 7.4: Markov stability analysis of $(T_2G_4)_4$ and two form 3 unimolecular quadruplexes containing fewer tetrads than their sequence suggests. When only a subset of the guanine bases are involved in the formation of tetrads, our methodology successfully distinguishes the guanine bases which are indeed part of the quadruplex tetrads core from those which only form part of a loop. Three tetrads are found for $(T_2G_4)_4$ and only two for human telomeric form 3.

also able to distinguish the bases actually participating in the formation of a tetrad from those which are not.

7.2.5 Comparison of quadruplexes with different number of strands

DNA quadruplexes formed by one, two or four separate strands (see Figure 7.1) largely differ in their physical properties such as thermal stability, kinetics, polymorphism, and sensitivity to the ion type and concentration. Tetramolecular structures, for instance, often have higher melting temperatures than bimolecular quadruplexes, which are themselves more stable than the unimolecular ones. Unimolecular quadruplexes, unlike quadruplexes formed by multiple strands, have been resolved in a wide range of different topologies between which they are thought to interconvert dynamically (Lee *et al.*, 2005). Ion type and concentration also have a strong impact on their topologies and stability. Conversely, bimolecular quadruplexes such as the

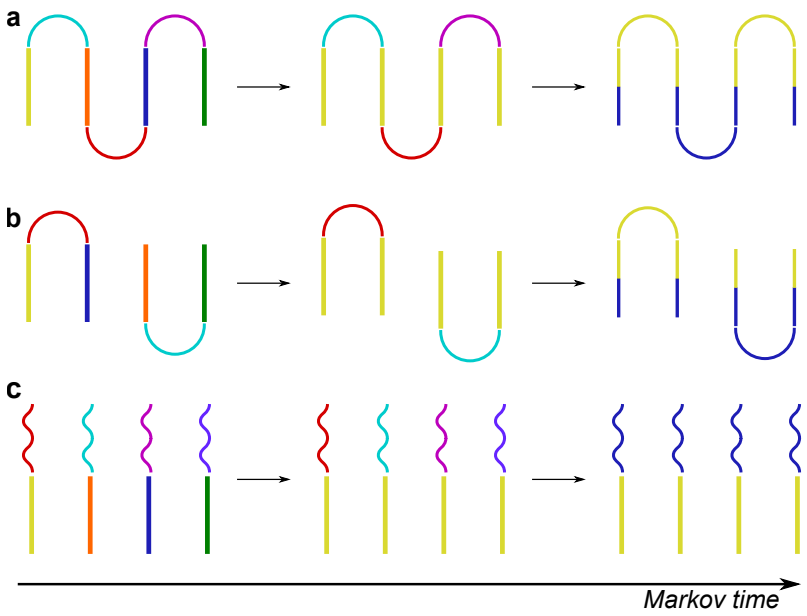


Figure 7.5: Schematic "unrolled" view of the multiscale partitioning pathway of most uni- (**a.**), bi- (**b.**) and tetramolecular (**c.**) DNA quadruplexes analysed here. Loops (curved lines) are mostly separated from the guanine bases involved in the tetrads (straight lines) and only become integrated into global interstrand communities at the largest scales, where the quadruplexes are usually partitioned along one of the tetrads.

classic O. Nova d($G_4T_4G_4$) usually take a unique conformation and the ion type or concentration only have a minor impact on loop mobility (Hud and Plavec, 2006).

On Figure 7.5, we show in a schematic way how the multiscale organisation of the quadruplexes unfolds at the larger scales. The communities initially group bases locally along the sequence, and then evolve towards global clusters that associate bases from all strands, starting from the identification of the tetrads community, and until the division of the quadruplex into two along one of the G-tetrads and perpendicularly to the ion channel.

Although all quadruplexes share these broad characteristics, differences appear at the transitions between these main levels of organisation (see Figure 7.6). Firstly, bi- and tetramolecular quadruplexes exhibit persistent intermediary groupings taking place across pairs of strands, unlike unimolecular structures which only form interstrand communities at larger scales when the tetrads community, which joins bases from all four strands, is identified. Interestingly, the two strands that are joined together in the communities of the bimolecular quadruplexes also always belong to two different DNA molecules. Our identification of robust communities containing pairs of strands in bi- and tetramolecular quadruplexes could be indica-

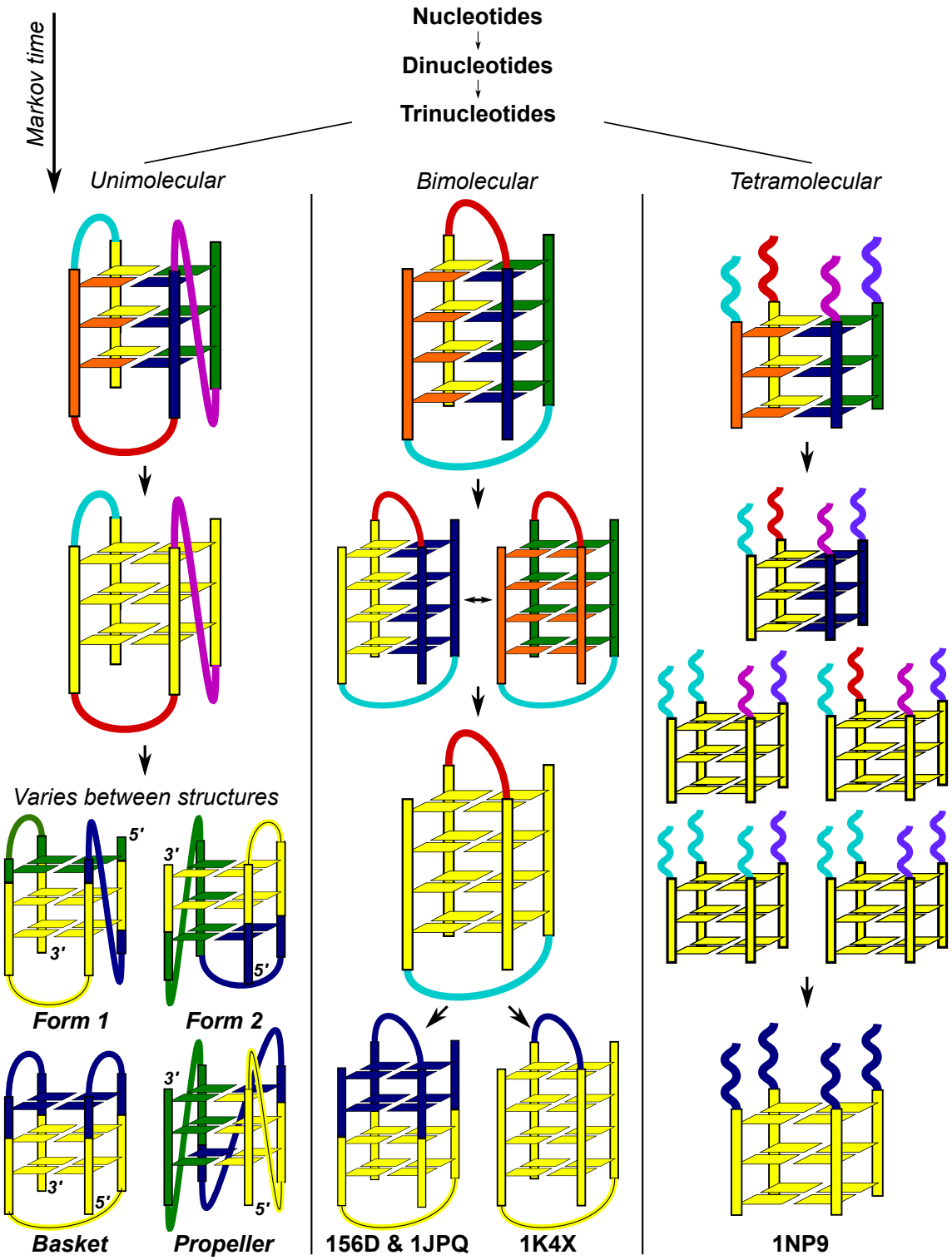


Figure 7.6: Overview of the multiscale partitioning of mono-, bi-, and tetramolecular DNA quadruplexes. All forms share a number of similar communities, such as the ones formed by the tetrads, the loops and the guanine bases of each individual strand, but differ in their intermediary groupings and last (coarsest) partitioning.

tive of the existence of stable intermediates of DNA duplexes that could be stabilised by base pairings between guanines along the association/dissociation pathway in bi- and tetramolecular structures. Interestingly, recent NMR experiments (Ceru *et al.*, 2014) published while this work was being written identified a new folding intermediate in the O. Nova d($G_4T_4G_4$) quadruplex where all guanine bases are involved in G-G base pairings as predicted by our communities. Tetramolecular quadruplexes have also been found to assemble in a step-wise manner via the formation of stable intermediate states which include DNA strand dimers and triplexes (Tran *et al.*, 2013).

Secondly, the distinction between loops and tetrads is less marked in tetramolecular structures than for the other quadruplexes. In tetramolecular quadruplexes, interstrand groupings are formed equally between both loop and tetrad residues, although G-tetrads still associate faster than loops. Compared to other quadruplexes, loops from different strands are thus clustered together earlier which suggests that they might be interacting more strongly in tetramolecular structures. These increased loop interactions could provide an extra contribution to the quadruplex physical stability, in addition to that given by the tetrads, which could explain the higher melting temperature which has been observed experimentally for tetramolecular quadruplexes.

Thirdly, the highly polymorphic unimolecular structures display a wider variety of partitions. At the largest scales, while most divide into two along one of the tetrads which results in mixed groups of loop and tetrad residues, double chain reversal loops generally remain more disconnected and form a well separated community up until the latest partition (notice the isolation of the double chain reversal loops in the propeller, form 1 and 2 unimolecular structures in Figure 7.6 as opposed to the basket-type). This behaviour is particularly marked in the propeller quadruplex which, unlike all the other structures, partitions exclusively along the sequence at the largest scales. This suggests that double chain reversal loops may be more stabilised by the local interactions between neighbouring nucleotides (such as pi-stacking and covalent bonds), than through global interactions between different strands (governed by hydrogen bonding and cation coordination). In turn, relatively weaker interstrand interactions relax the constraints imposed on the motions of each individual strand. This could in turn favour the high polymorphism observed in unimolecular quadruplexes. In the unfolding pathway of quadruplexes, our results suggest that strands connected through double chain reversal loops should be the most prone to dissociate from the quadruplex tetrads core.

Finally, with the exception of the propeller quadruplex, the tetrads community is always much more persistent over Markov times in bimolecular than in unimolecular structures. For tetramolecular structures, the tetrads community always corresponds to the final bipartition obtained at the largest scales, and thus has an infinitely long Markov lifetime. A higher persistence of this community is indicative of a higher propensity of the quadruplex to exhibit a global interstrand collective behaviour relative to local intrastrand interactions along the sequence. This could once again be a contributing factor to the physical stability as our results agree with the melting temperatures being generally the highest for tetramolecular quadruplexes and the lowest for unimolecular quadruplexes.

7.2.6 The community structure of the G quadruplexes predicts the folding/unfolding pathway

While early works focused on the equilibrium conformations, quadruplexes are now increasingly being studied from a dynamical perspective. As detailed in the introduction, some have indeed been found to exhibit a high polymorphism and to interconvert between different stable conformations in solution. It has now become clear that any static structure must be viewed in the broader context of these dynamical interconversions between multiple relevant conformational substates. Here, we use our methodology to shed light on the relation between the quadruplex structures and their dynamical behaviour. In particular, we relate their structural organisation as identified by our methodology to their unfolding process, comparing our results with the model developed by Ambrus and coworkers (Ambrus *et al.*, 2006; Zhang *et al.*, 2010) from NMR data, and the no-salt molecular dynamics simulations from Stadlbauer *et al.* (2013).

Due to the slow rates of interconversion between the conformational substates and the computational limitations of molecular dynamics simulations, Stadlbauer *et al.* (2013) chose to monitor the unfolding process of DNA quadruplexes under no-salt conditions[4] to make it observable under the time scale limitations of fully atomic molecular dynamics. Specifically, by running simulations of uni-, bi- and tetramolecular DNA quadruplexes starting from an experimental crystal or NMR structure from which they removed the cations, they obtained mechanistic insights into the unfolding process of quadruplexes. Similarly, we here study the impact of

[4]Since the coordinating cations are necessary to the formation and stability of the quadruplexes, molecular dynamics simulations conducted in a solution devoid of ions result in a rapid unfolding of the quadruplexes.

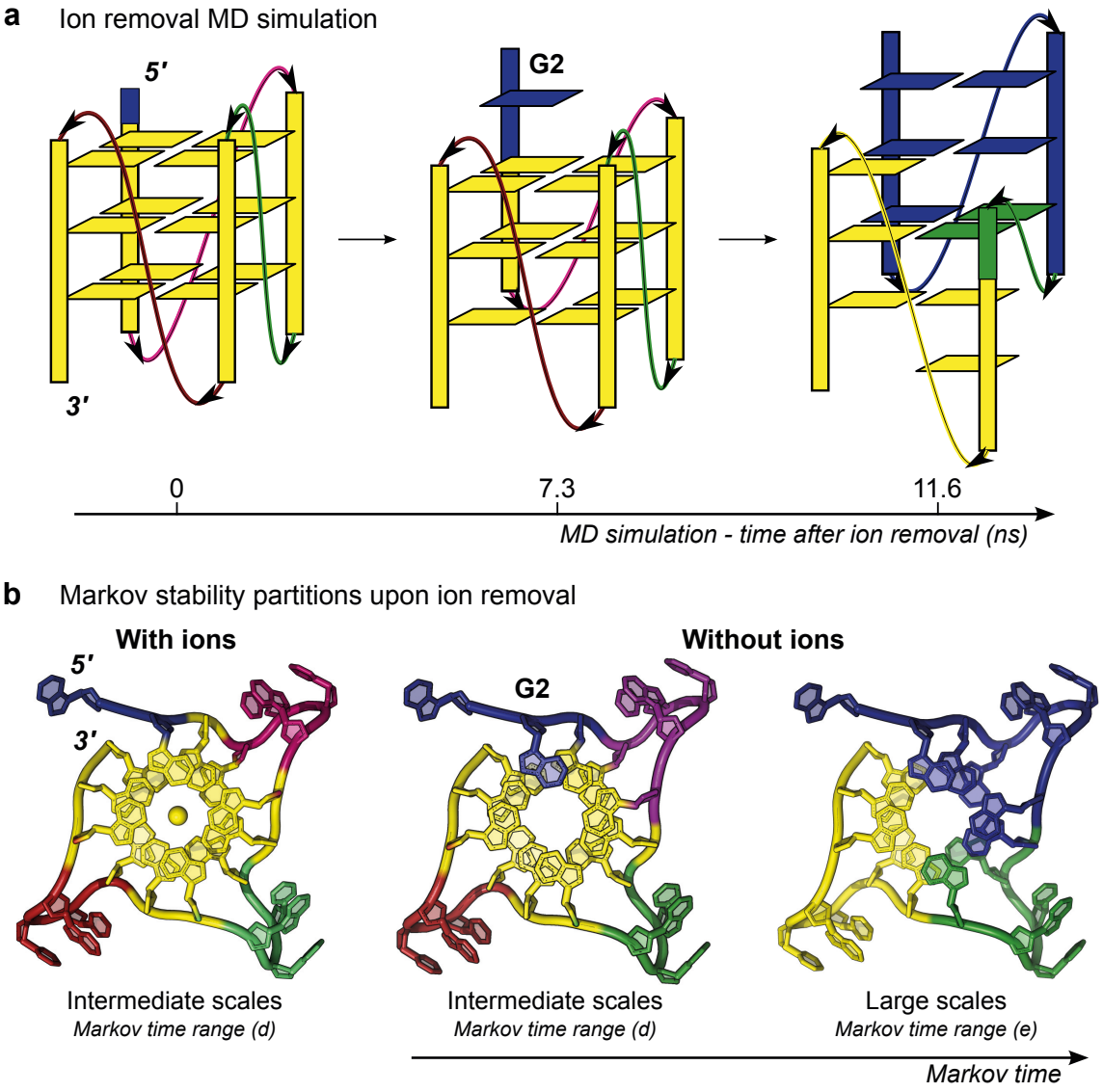


Figure 7.7: Bases belonging to a same community tend to remain together during the unfolding process. **a.** The three conformations correspond to snapshots of the unfolding process of the propeller type quadruplex at three different time steps (initial structure, and after 7.3 and 11.6 ns) as observed in no-salt molecular dynamics simulations conducted by Stadlbauer *et al.* (2013). The colouring of bases and loops is done according to our results (shown in **b.**) obtained using Markov stability on the graphs of the fully folded quadruplex NMR structure, with (simulation time 0) and without (simulation times 7.3 and 11.6 ns) the inclusion of the stabilising central cations. The colouring at 7.3 ns and 11.6 ns correspond respectively to the partitions identified in the range of Markov time (d) and (e) (see Figure 7.2) in the protein graph without the inclusion of the central cations. **b.** The partitions obtained using Markov stability shown here on the structures.

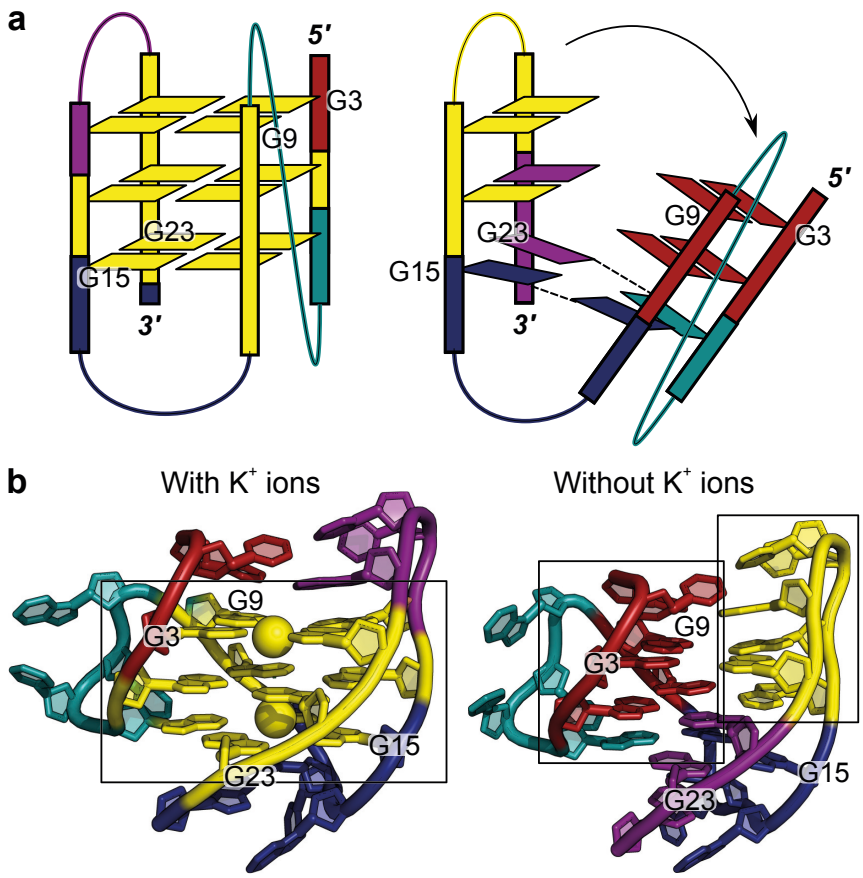


Figure 7.8: In form 1 quadruplexes, the bases grouped in a single community by Markov stability after the removal of the ions (**b**) match the bases moving together during molecular dynamics simulation of the unfolding process (**a**) (Stadlbauer *et al.*, 2013). The dashed lines indicate the presence of hydrogen bonds, and the colouring in **a** is done according to our results in **b**.

removing the cations on the structural organisation of each quadruplex, and relate our results to the unfolding pathway observed by Stadlbauer *et al.* (2013).

Our analysis of the multiscale structural organisation of the quadruplexes after removal of the stabilising cations shows that bases grouped together in the communities found by Markov stability tend to remain together in the early stages of the unfolding process. Figure 7.7a shows the first three relevant intermediate conformational states identified by Stadlbauer *et al.* under the no-salt molecular dynamics simulation of the propeller type human telomeric quadruplex (1KF1, see Figure 7.1). The unfolding proceeds via the slippage of the guanine base 2, followed by a further displacement resulting in the two strands from the 5' end separating from the last two 3' end strands. Figure 7.7b highlights the major changes in the partitions obtained by Markov stability upon the removal of the cations. Interest-

ingly, removing the cations maintains the tetrads community (although it strongly impacts its robustness, see Figure 7.2), with the exception of guanine 2 which is now dissociated from the tetrads community and grouped with the 5' terminal loop. The latest robust partition we identified divides the quadruplex into three groups (the last two 3' end strands, the first two 5' end strands and the middle loop) which appear to match closely the regions of the quadruplex that are kept together in the later stages of the unfolding pathway. In Figure 7.8a, the unfolding pathway of the form 1 unimolecular quadruplex, as observed in the same no-salt molecular dynamics simulations, is shown to proceed via an opening motion which keeps the 5' first two strands and 3' last two strands grouped together. Interestingly, removing the cations from the original structure also appears to disrupt their all-scale structural organisation (Figure 7.8b). In particular, the tetrads community disappears in favour of duplexes of strands corresponding to those moving together during the unfolding process.

To explain the interconversions they observed in their NMR data upon the addition of K^+ in a Na^+ solution, Zhang *et al.* (Zhang *et al.*, 2010; Ambrus *et al.*, 2006) proposed a pathway for a conformational transition between the basket type quadruplex and the hybrid forms 1 and 2, with hybrid form 3 as a stable intermediate. Their model (see Figure 7.9a) assumes the following steps: firstly, a slippage of one of the guanine bases occurs at the 5' end, giving rise to a transient 2-tetrad intermediate equivalent to form 3; then, the 5' end strand dissociates from the tetrads core, producing a triplex intermediate, and swings back to the other side of the second strand which results in a double chain reversal loop and the formation of the hybrid form 1 quadruplex. They further hypothesised that, considering the slow rate of interconversion between forms 1 and 2, form 2 unimolecular quadruplex could be generated through a symmetric pathway that would start from the 3' end instead of the 5' end.

Interestingly, the community structures found appear to correctly identify the dissociating strands suggested by Zhang *et al.* (2010). Markov stability indeed consistently identifies the 5' end strand in the basket type, and both the 3' and 5' end strands in form 1 and 2, as being always more disconnected from the tetrads core than the other two or three strands (see Figure 7.9b). At the Markov time where the G-tetrads community is identified, the backbone atoms from the 5' end (as well as the 3' end for forms 1 and 2) either form an independent community or become grouped with the neighbouring loop.

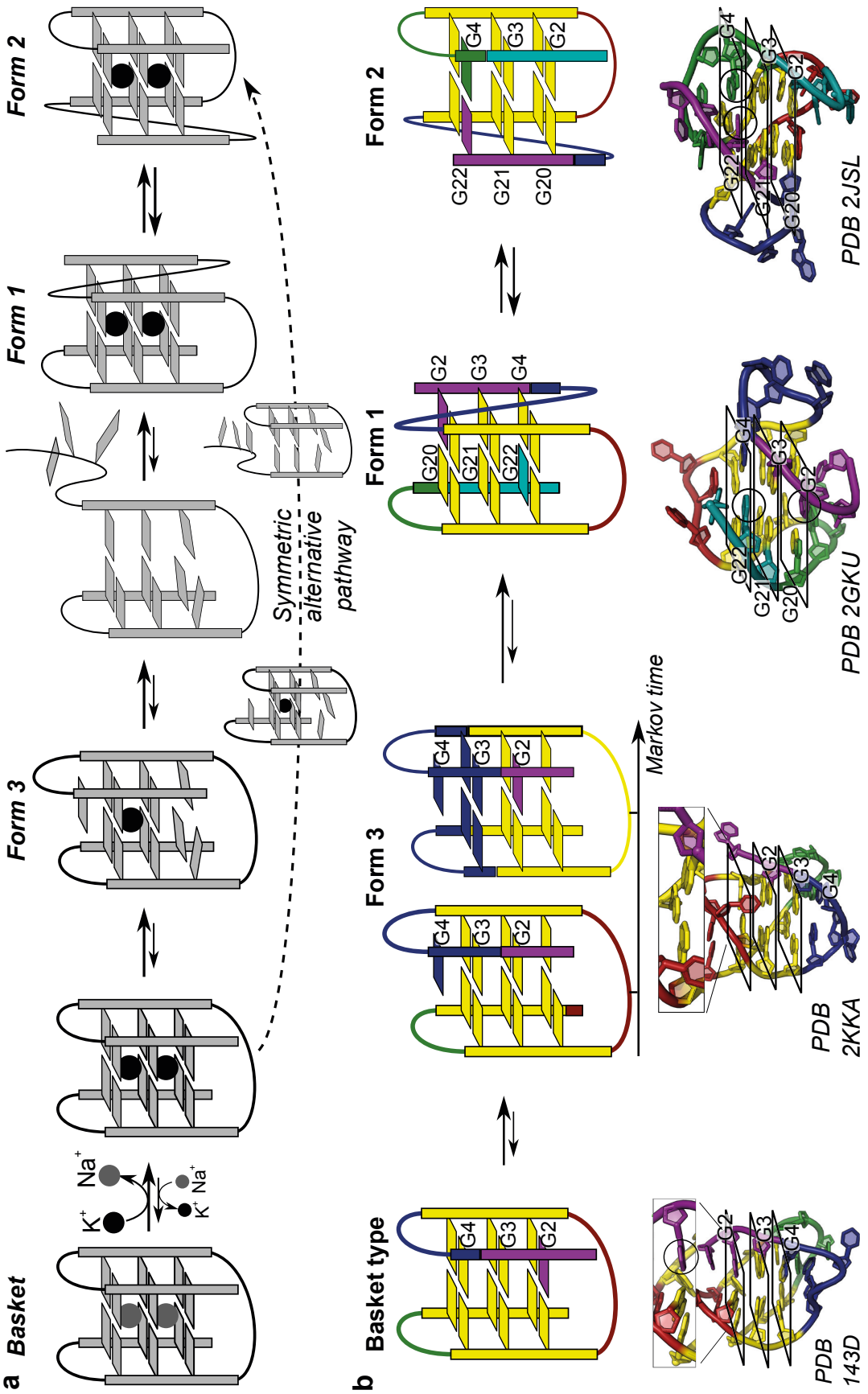


Figure 7.9: (Figure caption on the next page)

Figure 7.9: (Figure on the previous page) The partitions identified by Markov stability contain signatures of the interconversion pathway between the basket type, form 1 and form 2 unimolecular quadruplexes. For each topology, the strand showing a more pronounced disconnectivity in our analysis corresponds to the one that dissociate in the interconversion pathway. **a.** Schematic diagram of the interconversion pathway proposed by Zhang *et al.* (Ambrus *et al.*, 2006; Zhang *et al.*, 2010) between the Na^+ basket type and the hybrid form 1 and 2 telomeric quadruplexes. **b.** Relevant partitions obtained through Markov stability for the basket type, form 1, 2, and 3 DNA quadruplexes.

In addition, guanine 2 is entirely absent from the tetrads community in the basket type quadruplex suggesting that it is already primed for the unfolding in the fully folded structure. Similarly, in form 1 and 2, the last guanine base from the dissociating strand (5' end in form 1 and 3' end in form 2) is absent from the tetrads community. Finally, form 1 and 2 have both the 3' and 5' ends backbone atoms disjoined from the tetrads core.

Altogether, these results are in agreement with the interconversion pathway proposed by Zhang *et al.* (Ambrus *et al.*, 2006; Zhang *et al.*, 2010), but suggest that the hypothetical symmetric pathway which reaches form 2 directly through the slippage of the 3' end of the basket type should be much less prevalent, if it indeed exists. Instead, our results suggest that the preferred interconversion pathway towards form 2 is a direct route via the dissociation of both the 3' and the 5' strands from form 1, either each in turn through triplexes and an intermediary chair conformation, or simultaneously through an transient hairpin.

7.3 Discussion

In this chapter, we have shown that, using a detailed graph theoretical model adapted for nucleic acid molecules, the all-scale analysis of G-quadruplex structures recovers the elementary biochemical building blocks, but also relates a number of experimentally observed dynamical properties to the static structure of the quadruplexes. This allowed, for instance, to help establish the unfolding pathway or identify important structural elements for their physical stability.

In particular, we showed that the importance of ions in the formation and stability of the quadruplexes can be partly rationalised in terms of their impact, via their coordination of the guanine bases, on the existence and robustness of a cohesive substructure formed by the ensemble of guanine tetrads. The method itself was also

able to distinguish the guanine bases actively involved in the tetrads through hydrogen bonds, cation coordination and stacking interactions, from those effectively associated with the connecting loops. This could be used to identify guanine bases with a higher mobility, likely to dissociate from the tetrads core, and which should be targeted in priority in the design of stabilising drug compounds.

Double chain reversal loops emerged from our analysis with a clearly distinct behaviour from diagonal and lateral loops. Unlike the latter, double chain reversal loops always remain highly disconnected from the rest of the structure in all the quadruplexes analysed. They emerge as more independent from the tetrads core than other types of loops, and could therefore notably be expected to display a higher mobility. Their segregation is particularly pronounced in the propeller structure where, unlike any of the other quadruplexes analysed, the structural partitioning at the largest scale dissociates the tetrads core between each individual strand and the large scale dynamics is thus likely to be dominated by the loops. Interestingly, many of its properties have been found to be very different from the other unimolecular topologies (Lane *et al.*, 2008) and it has been hypothesised that the propeller form should not be found in normal solution, and only form under particular conditions such as those found in crowded or solution depleted environments, e.g. crystals or the cell nucleus (Heddi and Phan, 2011; Yu *et al.*, 2012). Its relevance *in vivo* has also been debated (Marchand *et al.*, 2013). Our results suggest that a possible higher mobility of the double chain reversal loops could be responsible for the destabilisation of the propeller quadruplex in solution. Loop motions could however be restricted by molecular crowding which would then stabilise the structure (Marchand *et al.*, 2013). In light of these observations, we suggest that double chain reversal loops might form a preferential target for stabilising compounds and we believe that their dynamics should be further investigated experimentally. Recent molecular dynamics simulations seem to agree with our general conclusions (Zhu *et al.*, 2013).

Our results also provide a possible link between the structural organisation of the quadruplexes and their thermal stability and polymorphysm. In our analysis, a higher propensity for the nucleotides to associate with bases from other strands, as opposed to their neighbouring bases along the sequence, appears beneficial for thermal stability and unfavourable for structural polymorphism. The expectation is that stronger interstrand associations could limit the ability of the individual strands to dissociate from the tetrads core to unfold or form a new conformation. In particular, the highly polymorphic unimolecular quadruplexes were the most likely to exhibit groupings along the sequence, while the very stable tetramolecular

quadruplexes exhibited additional interstrand groupings between loops which could further stabilise the quadruplex in addition to the tetrads. Finally, we found both bi- and tetramolecular quadruplexes to generally have a more robust community of G-tetrads, which could also contribute to their higher thermal stability (Neidle and Balasubramanian, 2006).

Finally, our analysis demonstrates that signatures of the quadruplexes unfolding pathway are already encoded in their structure, and can be successfully captured by our methodology. In particular, bi- and tetramolecular quadruplexes showed robust intermediary groupings of DNA strand duplexes. This suggests the possibility of stable bi-stranded intermediates along the association or dissociation pathways. This hypothesis is supported by previous experimental data on the kinetics of association of DNA tetramers (Wyatt *et al.*, 1996) suggesting the association of two dimers to form the tetramers as the rate-limiting step. In addition, the strands that appeared the most disconnected from the tetrads core in our analyses were those predicted to initiate the unfolding process in hypothetical pathways derived from both experimental data and molecular dynamics simulations. A comparison of no-salt molecular dynamics simulations with the Markov stability analysis of the same quadruplexes in the absence of the central cations showed that some of the main transition steps along the unfolding pathway could be predicted from the partitions obtained through Markov stability alone. The model of interconversion between the basket-type, form 1, form 2 and form 3 telomeric quadruplexes proposed by Zhang *et al.* (2010) was similarly consistent with the partitions we identified for each structure. We find that the regions which were more disconnected from the tetrads core in our analysis correspond to the first bases to dissociate during the unfolding or interconversion pathway proposed in their model. Our results also suggests the symmetric pathway speculated by Zhang *et al.* (2010), which yields the hybrid 2 form from the basket type quadruplex, to be much less prevalent than the direct conversion of form 1 into form 2.

Together, our results suggest that the dynamical behaviour of the DNA quadruplexes is, to a large extend, already encoded in their structure. As the dynamics of quadruplexes is paramount to the understanding of their biological role and the design of specific drugs, it is encouraging that some of its key aspects can be understood from the structural data alone. Considering the agreement of our results with experimental data and other computational analyses, our methodology could thus prove to be a powerful tool to predict the folding/unfolding pathway of other DNA

quadruplexes, and identify bases that should be preferentially targeted by stabilising drugs.

Despite the highly dynamical nature of the DNA quadruplexes, we here showed the ability of our computational framework to uncover significant structural features that relate to key dynamical properties, with a comparable accuracy to our analyses of the much more structured proteins.

Chapter 8

Conclusions

Decoding the internal machinery of biomolecular structures remains a true scientific challenge today. At its very heart lies the complexity emanating from the wide spectrum of time and spacial scales over which their structure and dynamics unfolds. This deep hierarchy of scales is inherent to protein functionality: the emergence of functional motions at large scales is the consequence of the individual contributions of the dynamics at the atomic level.

While the multiscale organisation is often assumed a priori, we have, in this thesis, taken the opposite view, and sought to unravel the multiscale organisation of biomolecules from the atomic structure as a means to improve our understanding of the mechanisms behind biological function. To this end, we introduced a graph theoretical framework for protein and nucleic acid structures inspired by biochemical force fields and multiscale community detection on networks. Acting as a computational microscope, it provides a lens to explore the structural "anatomy" of biomolecules at all scales, from atoms to the quaternary structure, by sweeping its focus. As such it seamlessly links the dynamics at different scales, without initial assumptions or the use of any a priori information other than the atomic interactions given by the structure and force fields.

In Chapter 3, we proposed a fully atomic energy-based network model of biomolecular structures that takes full account of the physico-chemical details of the atomic interactions along with their energies. As such, it contrasts with the common use of an unweighted network of residues based on distance cutoffs by providing a physically more realistic representation of the atomic interactions. Building on the earlier works of Delvenne *et al.* (2010) and Meliga (2009), and noting that any dynamics taking place on a network is shaped by the structure on which it unfolds, we pro-

posed the use of Markov stability (Delvenne *et al.*, 2010) to explore the structural organisation of biomolecules at all scales. The time evolution of the random process evolving on the graph was used as a way to reveal substructures that are relevant over particular time spans of the dynamics, and the robustness of our solutions, measured by the variation of information, to evaluate their biochemical significance.

The suitability of the methodology was shown in Chapter 4 on adenylate kinase, a simple enzyme whose structure, dynamics and functional mechanisms are well characterised. Firstly, we recovered the main biochemical building blocks including atoms, chemical groups, residues, secondary structure and functional domains. Secondly, we obtained a good agreement with diverse experimental data including hinge analyses, normal modes and open-closed conformational changes. Thirdly, we designed two types of biochemically motivated surrogate random graph models which showed that biochemically meaningful substructures were found with high significance, and allowed us to correctly identify the scale at which particular types of interactions influence the structure and dynamics of the molecule. Finally, our analyses of the open and closed forms of AdK revealed an almost unchanged structural organisation between the two highly distinct conformations. This remarkable property of AdK structure could be linked to its ability to spontaneously sample the closed conformation in the absence of substrate.

In Chapter 5, our methodology was further expanded and used towards improving our understanding of MTIP, a small yet largely unexplored myosin light chain. Long-lived clusters identified were found to correspond to regions with well defined dynamical properties, such as domains and rigid clusters, or with a distinct functional role. Several of our conclusions were subsequently verified and further explained by experiments, including the loss of physical stability in the unbound state which was associated with less robust community structures, the destabilisation of the N-terminal end α-helix which we found to be highly dissociated from the rest of the structure, and the very distinct dynamics undergone by the C-terminal end which appeared in our analysis as the most disconnected region of the C-terminal domain. Taking advantage of the low computational cost of our approach and its all-scale property allowing evaluate the impact of individual bonds on the largest scales, a computational alanine scanning mutagenesis tool was developed which successfully identified five of the six MyoA tail residues experimentally found to be critical for binding, and suggested a seventh yet to be verified in binding assays.

The contribution of our approach also lies in its ability to deal with very large structures with rich and complex dynamics without the need to sacrifice the level of

detail in the model. This property was fully exploited in Chapter 6, where Markov stability was used to study the workings of multimeric proteins characterised by collective quaternary events that are initiated by local atomic changes propagating bottom-up through the secondary and tertiary levels. The methodology was further expanded by taking into account the ensemble of suboptimal yet meaningful solutions which, plotted as heatmaps, provided a visualisation of the underlying landscape of the biologically relevant partitionings. Our analysis of Rubisco revealed, at the level of the barrel domain, functional units and quaternary structure, the coexistence of multiple structural organisations dominating in alternance at different stages of the catalysis. Two symmetric partitions, in particular, involved a sharing of the small subunits at the last stages of the reaction and upon inhibition, and suggested the existence a communication channel accross subunits which could be linked to forms of cooperative mechanisms. In ATCase, a classic example of cooperative enzyme, a similar behaviour was observed which disappeared in the presence of CTP, an effector found to negatively impact cooperativity. Contrasting our results with hemoglobin, a globular structure, suggested that biomolecules with highly complex functional mechanisms tend to display a much richer structural organisation across scales.

In Chapter 7, the methodology was generalised to nucleic acid structures, which present a real challenge for most computational methods. Unlike proteins, DNA molecules are generally less structured, more flexible and sample a large variety of configurations *in vivo*. Focusing on G-quadruplexes, four-stranded DNA structures held together by planar quartets of guanine bases, our analysis established a link between their dynamical and polymorphic properties and characteristics of their structure. Double-chain reversal loops appeared as highly disconnected from the tetrads core, suggesting a higher mobility that could favour polymorphism. Analysing the robustness of communities encompassing different strands, which was found to be a good predictor for the physical stability of the structures, highlighted the stabilising role of the central cations. Finally, nucleotides that appeared as more disconnected from the structure in our analysis were shown to identify the first bases to dissociate from the tetrads core and initiate the unfolding process in two models for quadruplexes folding/unfolding pathways.

8.1 Future work

The present work opens several research directions, including further experimental and computational analyses of the protein and DNA molecules investigated here, as well as new methodological developments and other applications.

8.1.1 Further analyses of the biomolecules studied

Hypotheses derived from our analyses of the different protein and DNA structures should be verified experimentally. In the case of MTIP, the effective impact on the binding affinity of the MyoA tail alanine 809 detected by our mutational analysis could be assessed in new binding assays. Further testing for cooperativity in Rubisco with mutated L-S interface residues could help verify and better characterise the communication patterns that take place between the different subunits. Concerning the G-quadruplexes, NMR or FRET could be used to further investigate the dynamics of double chain reversal loops, which our analysis suggests to have a destabilising effect, along with mutations or compounds that specifically target them. The bases we found to be more disconnected from the tetrads core could similarly form preferential targets for drugs aimed at stabilising the structures.

While we studied the changes in the structural organisation of spinach Rubisco throughout the catalysis, applying the same computational analysis across multiple species would likely provide a deeper insight into the relation between structure, specificity and kinetics in Rubisco. Similarly, the role of particular residues could be further evaluated in mutant structures or by estimating their impact on the community structure at different scales. Such analyses could help suggest new mutagenesis studies to uncover the structural rules that regulate Rubisco's activity by enhancing or blocking particular communication pathways in the structure. RNA quadruplexes, which have recently been shown to form *in vivo* (Biffi *et al.*, 2014), would be a valuable comparison to our analysis of DNA quadruplexes. They exhibit distinct properties of self-assembly, stability and ligand binding to their DNA counterpart, and are involved in a wide range of biological processes, notably through their probable presence in many mRNAs (Collie and Parkinson, 2011).

8.1.2 Methodological developments

Although water molecules have not been included in the graphs, the effect of solvent has been implicitely taken into account through the use of a hydrophobic potential

of mean force as well as criteria based on distance and atom types. Water is however not an inert medium, but plays an active role that fundamentally impacts the structure, dynamics and function of biomolecules. Water molecules can link different parts of the biomolecule through water-mediated hydrogen bonds, form ordered clusters on its surface, or be buried inside the core of the protein (Levy and Onuchic, 2006). Water also acts as a "lubricant", facilitating particular conformational changes by lowering energy barriers, and its explicit inclusion in some models is essential for protein folding prediciton (De Los Rios and Caldarelli, 2000; Papoian *et al.*, 2004). Explicitely including water molecules in our graph-theoretical framework is straightforward, and our approach could therefore be used to provide insights into the role of water on protein structure and dynamics.

Our definition of Markov stability assumes the stationary distribution of the random walkers (uniform in this case) as the initial state of the process. However, the perturbations are likely to be more important on the surface, in the active and allosteric sites than inside the biomolecule. The initial probabilities can however be readily adapted and studying the impact of different perturbations could provide a better insight into the behaviour of the protein or DNA molecule, and better model particular experimental setups. If more information on the time-evolution of the perturbation can be obtained, for instance through short simulations, an approximation of the actual diffusion process observed on the structure could also be directly used in place of the Markov process using a recently proposed temporal version of Markov stability (Petri and Expert, 2014).

Markov stability appears to establish a link between the Markov time at which a particular meaningful substructure is identified and the time scale of its associated dynamics, and a monotonic relation between the two has indeed already been observed before this work (Meliga, 2009). In spite of the general agreement of our results with experimental data, the derivation of a physical model for the Markov diffusion process is however lacking. As noted by Reuveni *et al.* (2010b,a), a mapping can be established between a random walk on a graph and the vibrational motion of the atoms in the protein, according to which our results could be reinterpreted in terms of the local trapping of atomic oscillations. However, this model requires the assumption of perfectly isotropic atomic motions which leads to unphysical characteristics of the potential (Thorpe, 2007) (see Section 2.2.2).

This work is also applicable to other problems in biochemistry. Allostery in particular, whereby the binding of a molecule to one site controls the activity of the protein taking place at a sometimes distant active site, ultimately relates to long-

range communication through the network of interacting residues. A wide range of graph-theoretical approaches have unsurprisingly been used with success (see Section 2.2.1) and, recent work by Amor *et al.* (2014) combining Markov stability with an analysis of the random walk transients was indeed used to reveal intra-molecular signaling pathways in caspase-1.

Due to the range of time scales spanned by the dynamics of biomolecules, coarse-graining is often a necessity for many simluation methods. Although systematic approaches exist (see for instance (Sinitskiy *et al.*, 2012)), coarse-graining is most often a choice of the modeller based on a priori information. Our partitioning method could therefore help provide a meaningful coarse-graining for simulations that is directly derived from the fully atomic structural information.

8.2 Final comments

The main contribution of this work has been to explore the functional mechanisms of biomolecules by relating their atomic structure to the dynamics over the entire spectrum of time and spatial scales, from bond vibration to domain displacement, from atoms to the quaternary structure. Its computationally inexpensive nature opens up time and length scales often inaccessible to traditional methods and, through the graph representation, information about the global structural organisation and slow dynamics of some of the largest proteins and protein assemblies could be extracted while keeping atomic details in the model.

Beyond the methodological developments and individual results for specific proteins or DNA quadruplexes, the intention has also been to offer a different perspective on the study of biomolecular structures. Traditionally, the analysis of protein structural data has been inclined to focus on either coarse global measures (e.g. root mean square deviation, solvent accessible area) or on the detailed study of local structural changes (e.g. rotation of individual side chains, large domain displacements). Similarly, computational methods have generaly considered the different scales rather independently, targeting for instance localised atomic vibrations (e.g. nanosecond molecular dynamics simulations), rigid regions spanning a few residues (e.g. combinatorial rigidigy), or the largest scale molecular motions that drive functional conformational transitions (e.g. analysis of the slowest normal modes). Yet the function of biomolecules is defined by the coupled interplay of their dynamics at all scales (Henzler-Wildman *et al.*, 2007a; Henzler-Wildman and Kern, 2007; Yaliraki and Barahona, 2007; Leitner *et al.*, 2006). Computational tools that can link

atomic alterations to their impact on the structural organisation at all scales open up a new perspective to study the intrinsic linkage between structure, dynamics and function in biomolecules.

Appendix A

Parameters for the construction of the graphs

In this appendix, we detail the parameters and potential energy functions used to identify and weight the edges of the graphs for protein and DNA structures.

Covalent bond

Covalent bonds are modelled by a single edge between the two atoms bound and are identified by FIRST from interatomic distances combined with a dictionnary of covalent bonds in standard amino acids. The weights given to covalent bonds correspond to the standard tabulated values of bond dissociation energies (Huheey *et al.*, 1993) (see Table A).

Bond	**Energy** (kJ/mol)	**Bond**	**Energy** (kJ/mol)	**Bond**	**Energy** (kJ/mol)
H — H	432	H — Se	276	C — O	358
H — C	411	C — C	346	C = O	799
H — N	386	C = C	602	C — S	272
H — P	322	C — N	305	P — O	335
H — O	459	C = N	615	P = O	544
H — S	363	C — P	264		

Table A.1: Energy used to weight the edges for covalent bonds in the graphs of proteins and DNA. Values correspond to the covalent bonds dissociation energy as given in reference (Huheey *et al.*, 1993).

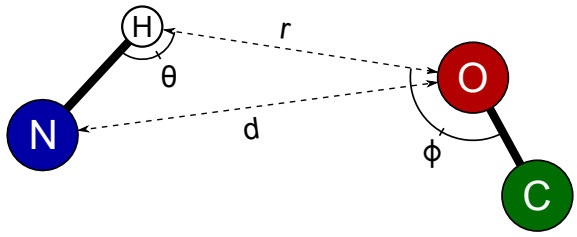


Figure A.1: Variables used to identify and compute the energy of covalent bonds and salt bridges.

Hydrogen bonds

Hydrogen bonds and salt bridges are assigned as an edge between a hydrogen and an acceptor atom if their distance is less than 2.6Å, the donor-acceptor distance is less than 3.6Å and the donor-hydrogen-acceptor angle is between 90 ° and 180 ° .

The enery weight associated with hydrogen bonds is given by the modified Mayo potential (Rader *et al.*, 2002; Dahiyat *et al.*, 1997):

$$E_{\mathrm{HB}} = V_0 \left[5 \left(\frac{R_0}{R} \right)^{12} - 6 \left(\frac{R_0}{R} \right)^{10} \right] F(\theta, \phi, \psi) \tag{A.1}$$

with

$V_0 = 8\mathrm{kcal/mol}$

$R_0 = 2.80\text{Å}$

$$F(\theta, \phi, \psi) = \begin{cases} \cos^2(\theta) \mathrm{e}^{-(\pi-\theta)^6} \cos^2(\phi - 109.5) & \text{for } sp^3 \text{ donor- } sp^3 \text{ acceptor} \\ \cos^2(\theta) \mathrm{e}^{-(\pi-\theta)^6} \cos^2(\phi) & \text{for } sp^3 \text{ donor- } sp^2 \text{ acceptor} \\ \cos^4(\theta) \mathrm{e}^{-2(\pi-\theta)^6} & \text{for } sp^2 \text{ donor- } sp^3 \text{ acceptor} \\ \cos^2(\theta) \mathrm{e}^{-(\pi-\theta)^6} \cos^2(\max[\phi, \psi]) & \text{for } sp^2 \text{ donor- } sp^2 \text{ acceptor} \end{cases}$$

where θ is the donor-hydrogen-acceptor angle, ϕ the hydrogen-acceptor-base angle, and ψ the angle between the normals to the planes defined by the 3 atoms around both sp2 centre.

For salt bridges, the edge weight is computed by (Dahiyat *et al.*, 1997)

$$E_{\mathrm{SB}} = V_0 \left[5 \left(\frac{R_s}{R + x} \right)^{12} - 6 \left(\frac{R_s}{R + x} \right)^{10} \right] \tag{A.2}$$

with

$V_0 = 8\text{kcal/mol}$
$R_0 = 3.2\text{Å}$
$x = 0.375\text{Å}$

The energy threshold for hydrogen bonds and salt bridges was set at 0.01 kcal/mol.

Hydrophobic tethers

Hydrophobic tethers are identified using the software package FIRST. Edges are placed only between pairs atoms if they are either carbon or suflur atoms, are within 8Å from each other, and are themselves covalently bound to a carbon, sulfur or hydrogen atom. In addition, only one hydrophobic tether edge is allowed per atom.

Hydrophobic tether edges are weighted using a simplified potential based on the hydrophobic potential of mean force proposed by Lin *et al.* (2007). Hydrophobic interactions are given an energy of -0.8 kcal/mol when the distance between the two atoms is below 5Å and -0.2 kcal/mol when it is below 8Å.

Π-stacking interactions

In proteins, edges for π-stacking interactions are assigned by FIRST using a 5.5Å distance cut-off and a 30 ° angle cut-off between the aromatic rings, as well as a 40 ° angle cut-off between the normal to the rings and the vector joining the ring centres. Edges are given a fixed weight of 10 kcal/mol corresponding to a typical energy of interaction (Sponer *et al.*, 2008).

In DNA structures, the weights are computed more accurately using the potential proposed by Hunter and Sanders (1990). The total energy is given by the sum of the contributions from van der Waals and electrostatic interactions:

$$E_{\text{stacking}} = \sum_{ij} \left[K_i K_j \left[C \exp(-\alpha \frac{r_{ij}}{r_{ij0}}) - \frac{A}{r_{ij}^6} + \sum_{kl} \frac{332}{\epsilon} \frac{q_i^k q_j^l}{r_{ij}^{kl}} \right] \right]. \qquad \text{(A.3)}$$

Here, the outer sum extends over all pairs of atoms belonging to two different aromatic rings. r_{ij} represents the distance between the two atoms, and K_i, K_j, C, A, r_{ij0} and α are standard parameters (Caillet and Claverie, 1975) (see Box 1). The

electrostatic contribution (Warshel *et al.*, 2006) is modelled by the third term. The sum extends over indices k and l which corresponds to three different point charges associated with each atom: one σ point charge at the nucleus of the atom and two π point charges placed 0.47Å above and below the aromatic plane.ϵ is the dielectric constant which we fixed at the most commonly used value of 4 (Gilson and Honig, 1986). The identification of π-stacking interactions in DNA does not rely on the geometric criteria set by FIRST. Instead, the total π-stacking energy of interactions is computed between every pair of bases and edges are assigned using an energy threshold set at 0.6 kCal/mol, corresponding to the energy of thermal fluctuation at room temperature.

The standard parameters used in Equation A.3 are as follows (Caillet and Claverie, 1975):

$$\alpha = 12.35 \quad A = 0.214 \quad C = 47 \times 10^3$$

$$r_{ij0} = \sqrt{\left(2r_i^W\right)\left(2r_j^W\right)}$$

where r_i^W is the van der Waals radii of atom i:

$$r_H^W = 1.2\text{Å} \quad R_C^W = 1.7\text{Å} \quad R_C^W = 1.77\text{Å} \quad R_N^W = 1.6\text{Å} \quad R_O^W = 1.5\text{Å}.$$

Finally, the parameters K_i also depend on the atomic specie:

$$K_H = 1 \quad K_C = 1 \quad K_N = 1.18 \quad K_O = 1.36.$$

Box 1: Parameters for the π-stacking interaction potential used for nucleic acids (Equation A.3).

Electrostatic interactions

Electrostatic interactions are generally neglected, unless they play a crucial role in the dynamics or function of the protein or DNA, such as the coordination with metal ions and the electrostatic interactions between the negatively charged phosphate groups of the DNA backbone.

Interactions with coordination ions are identified directly from the structure using the LINK entries from the PDB file (Bernstein *et al.*, 1978) and weighted using the Coulomb potential

$$E_{\text{coulomb}}(q_1, q_2, r_{12}) = \frac{332}{\epsilon}\frac{q_1 q_2}{r_{12}} \tag{A.4}$$

Adenine

Atom name	σ **charge**	π **charge**
N9	1.7057	-0.8623
C8	1.1687	-0.4342
H8	0.0675	0
N7	0.7074	-0.6358
C5	1.1684	-0.5253
C4	1.24	-0.4778
N3	0.7185	-0.6049
C2	1.2027	-0.4647
H2	0.0698	0
N1	0.722	-0.6106
C6	1.2726	-0.4446
N6	1.4443	-0.9399
H62	0.2076	0
H61	0.2076	0

Thymine

Atom name	σ **charge**	π **charge**
N1	1.6939	-0.8725
C6	1.0444	-0.4718
C5	1.009	-0.5817
C4	1.1808	-0.4286
N3	1.5617	-0.9048
C2	1.2671	-0.4194
O2	0.9448	-0.6752
O4	0.9329	-0.6635
C7	-0.104	0.0175
H6	0.059	0
H3	0.1914	0
H71	0.0409	0
H72	0.0409	0
H73	0.0409	0

Guanine

Atom name	σ **charge**	π **charge**
N9	1.7056	-0.8583
C8	1.1687	-0.4676
H8	0.0675	0
N7	0.7065	-0.6293
C5	1.1631	-0.5819
C4	1.2401	-0.474
N3	0.7249	-0.696
C2	1.363	-0.3942
N2	1.4534	-0.9387
H22	0.2087	0
H21	0.2087	0
N1	1.5729	-0.8615
H1	0.1926	0
C6	1.192	-0.4291
O6	0.9351	-0.6696

Cytosine

Atom name	σ **charge**	π **charge**
N1	1.6964	-0.8398
C2	1.2847	-0.4263
N3	0.7183	-0.6976
C4	1.2651	-0.4125
C5	0.9713	-0.5881
C6	1.0481	-0.4165
O2	0.9471	-0.6776
N4	1.4431	-0.9416
H5	0.0552	0.0175
DH6	0.0593	0
H42	0.2075	0
H41	0.2075	0

Table A.2: Parameters for the π-stacking interaction potential used for nucleic acids (Hunter and Sanders, 1990). We thank Christopher A Hunter for having kindly provided us with these parameters.

where r_{12} is the distance between the two atoms, q_1 and q_2 their two charges (in electronic unit), and ϵ is the dielectric constant ($\epsilon = 4$ in generally chosen in proteins (Gilson and Honig, 1986)). Charges for the residues were taken from the OPLS-AA force field (Jorgensen and Tirado-Rives, 1988) and for all non-standard residues and ligands, charges were obtained using the webserver of PRODRG (Schüttelkopf and van Aalten, 2004).

Similarly, the electrostatic repulsion between the negatively charged phosphate groups of the DNA backbone were included and weighted using the same Coulomb potential, with the addition of the Manning counterion and Debye screening effect (Swigon, 2009; Ravishanker *et al.*, 2007; Manning, 1978):

$$E_{\text{backbone}}(r_{12}) = \frac{332}{\epsilon}\frac{\delta^2}{r_{12}}\frac{\mathrm{e}^{-r_{12}}}{\lambda r_{12}} \tag{A.5}$$

where δ is charge partially neutralized by the counterion condensation effect, called the effective charge, and λ is the Debye screening length. For a monovalent salt such as NaCl, $\lambda = 3.0395\sqrt{c}$Å where c is the ionic concentration and $\delta = 0.24$.

Our computations however showed that the DNA backbone electrostatics had no impact on the final results our the Markov stability analyses. These interactions indeed mostly take place between adjacent residues which are only a few covalent bonds away from each other. The path joining adjacent bases in the graph is therefore dominated by the high-energy covalent bonds and the additional weak links created by the electrostatic interactions are negligible.

Appendix B

Robustness of the graph construction

The results of the Markov stability analysis are relatively insensitive to the exact value of the edge weights. As an exemple, we computed 100 Markov stability analyses of PfMTIP (PDB 2QAC, see Chapter 5) where, for each run, we added a gaussian random noise to the edge weights. The random perturbations were chosen with a zero mean and a standard deviation equal to 10% of the edge weights, such that

$$A_{ij}^{\text{pert}} = A_{ij}(1 + 0.1 \times \mathcal{N}(0, 1))$$

where $\mathcal{N}(0, 1)$ designates a normal random variable with zero mean and unitary variance, and A_{ij} is the weight of the edge between nodes i and j. The results of our analysis, shown in Figure B.1, indicate that the final partitions between the different randomized graphs is small and always well below the intrinsic variability of the solutions as measured from the Louvain initial conditions (see Chapter 3). The results of our analysis are thus almost unchanged after the modification of the edge weights. A precise weighting of the edges is consequently unnecessary for the purpose of this work, and the majority of the most commonly used force fields should yield identical results.

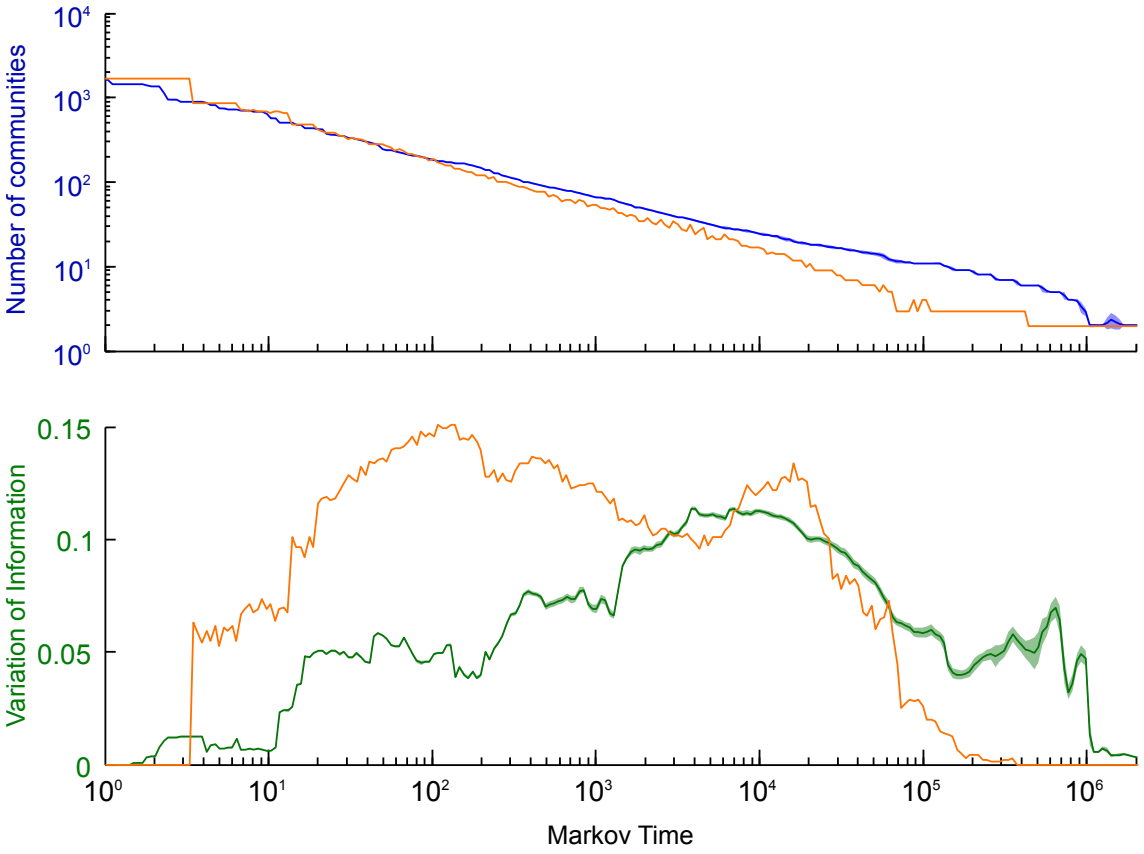


Figure B.1: The weighting of edges matters, but the final result is insensitive to small variations in edge weights. The plot shows the Markov stability analysis of 100 runs of PfMTIP (PDB 2QAC) with randomized edge weights. The blue and green curves indicate the mean of the number of communities and variation of information (based on the Louvain initial conditions) respectively. Shaded areas correspond to one standard deviation of the distribution at each Markov time. The orange curves corresponds to the number of communities and variation of information obtained for the equivalent unweighted graph (same edges, but with uniform weights).

Appendix C

Finding the optimal partition in practice

Once the graph is constructed from the protein structural data, we carry out the community analysis of this graph by optimising the Markov stability quality function (Delvenne *et al.*, 2010). The optimisation is carried out using the Louvain algorithm (Blondel *et al.*, 2008)—see Figure C.1 for an example. For each Markov time, the Louvain optimisation is repeated for several random initializations (usually 100 for small graphs and 1000 for larger graphs such as Rubisco) and the ensemble of solutions found is kept. We then report the optimal of all the solutions found at each time and we also calculate the mean variation of information (VI) of the ensemble of solutions obtained. The VI measures the dissimilarity of the optimised solutions found in the ensemble of runs of the Louvain algorithm, and thus serves as a description of the robustness of the optimal solution to the optimisation. As we sweep Markov time, the relevant partitions should be robust, i.e., they are found with high reproducibility by different initial randomisations of the Louvain optimisation. As the Markov time is increased, the quality function to be optimised (i.e., Markov stability) changes and the optimal partitions for each time will be different, i.e., as the Markov time changes, so does the ranking of the different partitions according to their Markov stability.

Figure C.1 shows the evolution of the Markov stability of five different partitions with the Markov time, and how the curve of the total number of communities emerges. The number of communities of the optimal partition (and the VI of the optimisation ensemble) for one of the structures are presented in Figure C.1a, while Figure C.1b shows the Markov stability of different partitions (in a range of Markov times) to indicate how the different partitions become optimal over different time intervals. Partitions which are optimal over a large range of Markov times and which

are also robust to the optimisation (as given by the VI) usually relate to well-defined, relevant levels of organisation.

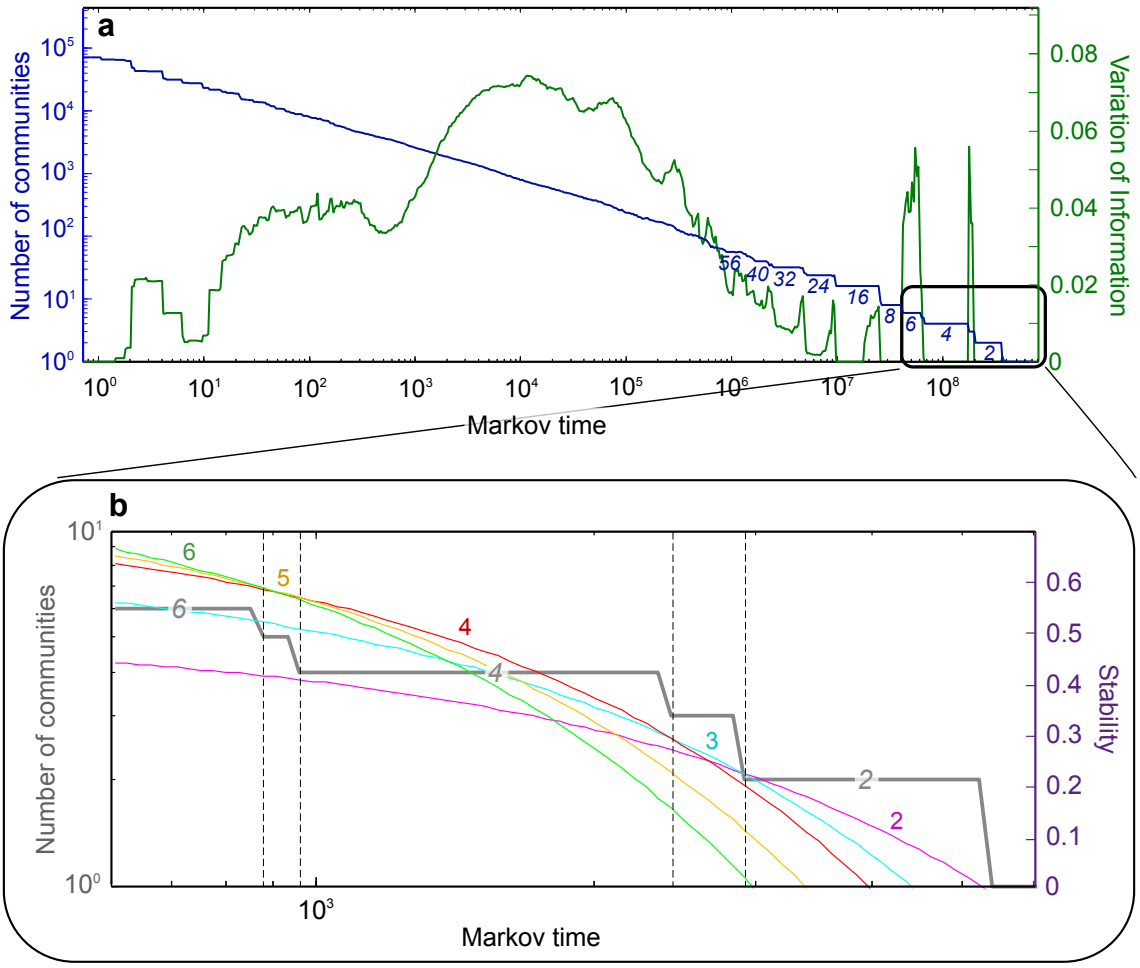


Figure C.1: (**a.**) Markov stability analysis of activated unliganded spinach Rubisco (1AUS) showing the number of communities (blue) and the variation of information (green) as a function of the Markov time. (**b.**) Zoom of the large Markov time region displaying the Markov stability for five partitions. As the Markov time increases, the Markov stability of the different partitions changes along with their ranking. This results in different partitions being optimal over different time intervals. Here, the partitions into 4 (in red) and 2 (in magenta) communities are optimal for a broad range of Markov times, while the partitions into 5 (orange) and 3 (cyan) communities are optimal only for small ranges of Markov times, indicating lower robustness of these partitions.

Appendix D

List of publications and publication permissions of third parties

Chapters 3, 4 and 5

The majority of the work in these chapters has been published in:

- Delmotte, A., Tate, E. W., Yaliraki, S. N., & Barahona, M. (2011). *Protein multi-scale organization through graph partitioning and robustness analysis: application to the myosin-myosin light chain interaction.* Physical Biology, **8**(5), 055010.

The email confirmation granting permission is reproduced below.

```
Dear Mr Delmotte,

Thank you for your request to reproduce IOP Publishing material.

We are happy to grant permission for the use you request on the terms set
out below.

If you have any questions, please feel free to contact our Permissions
team at permissions@iop.org.

I should be grateful if you would acknowledge receipt of this email.
```

Kind regards,

Laura Sharples
Rights & Permissions Officer
IOP Publishing Ltd
Temple Circus, Temple Way, Bristol BS1 6HG

T: +44 (0)117 930 1001
F: +44 (0)117 920 0997

What do scientific publishers do?

Conditions

Non-exclusive, non-transferrable, revocable, worldwide, permission to use the material in print and electronic form will be granted subject to the following conditions:

- Permission will be cancelled without notice if you fail to fulfil any of the conditions of this letter.

- You will reproduce the following prominently alongside the material:

  * the source of the material, including author, article title, title of journal, volume number, issue number (if relevant), page range (or first page if this is the only information available) and date of first publication. This information can be contained in a footnote or reference note; or

  * a link back to the article (via DOI); and

  * if practical and IN ALL CASES for works published under any of the Creative Commons licences the words ‘‘© IOP Publishing. Reproduced with permission. All rights reserved‘‘

- The material will not, without the express permission of the author(s), be used in any way which, in the opinion of IOP Publishing, could distort or alter the author(s)’ original intention(s) and meaning, be prejudicial to the honour or reputation of the author(s) and/or imply endorsement by

Chapter 6

The majority of the work in this chapter has been submitted with the following title:

- Delmotte, A., Reynolds, M.T., Woscholski, R., Barter, L.M.C., Yaliraki, S. N., & Barahona, M. (2011). *The all-scale structural anatomy of Rubisco throughout its catalytic reaction.*

Chapter 7

The majority of the work in this chapter is due to be submitted with the following provisional title:

- Delmotte, A., Reynolds, M.T., Vilar Compte,R., Yaliraki, S. N., & Barahona, M. (2011). *Predicting DNA quadruplexes folding processes through multiscale graph partitioning*

Bibliography

Adcock, S. A. and McCammon, J. A. (2006). Molecular dynamics: survey of methods for simulating the activity of proteins. *Chemical reviews*, 106(5):1589–1615.

Alberti, P., Bourdoncle, A., Saccà, B., Lacroix, L., and Mergny, J.-L. (2006). DNA nanomachines and nanostructures involving quadruplexes. *Organic & biomolecular chemistry*, 4(18):3383–91.

Amadei, A., Linssen, A. B., and Berendsen, H. J. (1993). Essential dynamics of proteins. *Proteins*, 17(4):412–25.

Ambrus, A., Chen, D., Dai, J., Bialis, T., Jones, R. a., and Yang, D. (2006). Human telomeric sequence forms a hybrid-type intramolecular G-quadruplex structure with mixed parallel/antiparallel strands in potassium solution. *Nucleic acids research*, 34(9):2723–35.

Amitai, G., Shemesh, A., Sitbon, E., Shklar, M., Netanely, D., Venger, I., and Pietrokovski, S. (2004). Network analysis of protein structures identifies functional residues. *Journal of molecular biology*, 344(4):1135–1146.

Amor, B., Yaliraki, S. N., Woscholski, R., and Barahona, M. (2014). Uncovering allosteric pathways in caspase-1 using Markov transient analysis and multiscale community detection. *Molecular bioSystems*, 10(8):2247–58.

Andersson, I. (1996). Large structures at high resolution: the 1.6 A crystal structure of spinach ribulose-1,5-bisphosphate carboxylase/oxygenase complexed with 2-carboxyarabinitol bisphosphate. *Journal of molecular biology*, 259(1):160–74.

Andersson, I. (2008). Catalysis and regulation in Rubisco. *Journal of experimental botany*, 59(7):1555–68.

Arenas, A., Díaz-Guilera, A., and Pérez-Vicente, C. (2006). Synchronization Reveals Topological Scales in Complex Networks. *Physical Review Letters*, 96(11):114102.

Arora, K. and Brooks III, C. L. (2007). Large-scale allosteric conformational transitions of adenylate kinase appear to involve a population-shift mechanism. *Proceedings of the National Academy of Sciences of the United States of America*, 104(47):18496–501.

Atilgan, A. R., Akan, P., and Baysal, C. (2004). Small-world communication of residues and significance for protein dynamics. *Biophysical journal*, 86(1):85–91.

Atilgan, A. R., Durell, S. R., Jernigan, R. L., Demirel, M. C., Keskin, O., and Bahar, I. (2001). Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical journal*, 80(1):505–15.

Ayton, G. S., Noid, W. G., and Voth, G. A. (2007). Multiscale modeling of biomolecular systems: in serial and in parallel. *Current opinion in structural biology*, 17(2):192–8.

Bahar, I., Atilgan, a. R., and Erman, B. (1997). Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding & design*, 2(3):173–81.

Bahar, I., Chennubhotla, C., and Erman, B. (2007). Reply to "Comment on elastic network models and proteins". *Physical Biology*, 4(1):64–65.

Bahar, I., Lezon, T. R., Yang, L.-W., and Eyal, E. (2010). Global dynamics of proteins: bridging between structure and function. *Annual review of biophysics*, 39:23–42.

Bahar, I. and Rader, A. (2005). Coarse-grained normal mode analysis in structural biology. *Current opinion in structural biology*, 15(5):586–92.

Bainbridge, G., Madgwick, P., Parmar, S., Mitchell, R., Paul, M., Pitts, J., Keys, A. J., and Parry, M. A. J. (1995). Engineering Rubisco to change its catalytic properties. *Journal of experimental botany*, 46(special issue):1269–1276.

Barahona, M. and Pecora, L. (2002). Synchronization in small-world systems. *Physical Review Letters*, 89(5):54101.

Baum, J., Richard, D., Healer, J., Rug, M., Krnajski, Z., Gilberger, T.-W., Green, J. L., Holder, A. a., and Cowman, A. F. (2006). A conserved molecular motor drives cell invasion and gliding motility across malaria life cycle stages and other apicomplexan parasites. *The Journal of biological chemistry*, 281(8):5197–208.

Ben-Hur, A., Elisseeff, A., and Guyon, I. (2002). A stability based method for discovering structure in clustered data. In *Pacific Symposium on Biocomputing.*, pages 6–17.

Bergman, L. W. (2002). Myosin A tail domain interacting protein (MTIP) localizes to the inner membrane complex of Plasmodium sporozoites. *Journal of Cell Science*, 116(1):39–49.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer Jr., E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1978). The protein data bank: A computer-based archival file for macromolecular structures. *Archives of Biochemistry and Biophysics*, 185(2):584–591.

Besteiro, S., Dubremetz, J.-F., and Lebrun, M. (2011). The moving junction of apicomplexan parasites: a key structure for invasion. *Cellular microbiology*, 13(6):797–805.

Biffi, G., Di Antonio, M., Tannahill, D., and Balasubramanian, S. (2014). Visualization and selective chemical targeting of RNA G-quadruplex structures in the cytoplasm of human cells. *Nature chemistry*, 6(1):75–80.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., and Hwang, D. (2006). Complex networks: Structure and dynamics. *Physics Reports*, 424(4-5):175–308.

Bochman, M. L., Paeschke, K., and Zakian, V. a. (2012). DNA secondary structures: stability and function of G-quadruplex structures. *Nature reviews. Genetics*, 13(11):770–80.

Böde, C., Kovács, I. A., Szalay, M. S., Palotai, R., Korcsmáros, T., and Csermely, P. (2007). Network analysis of protein dynamics. *FEBS letters*, 581(15):2776–2782.

Bogan, A. A. and Thorn, K. S. (1998). Anatomy of hot spots in protein interfaces. *Journal of molecular biology*, 280(1):1–9.

Bongini, L., Piazza, F., Casetti, L., and De Los Rios, P. (2010). Vibrational entropy and the structural organization of proteins. *The European physical journal. E, Soft matter*, 33(1):89–96.

Bončina, M., Lah, J., Prislan, I., and Vesnaver, G. (2012). Energetic basis of human telomeric DNA folding into G-quadruplex structures. *Journal of the American Chemical Society*, 134(23):9657–9663.

Bosch, J., Turley, S., Daly, T. M., Bogh, S. M., Villasmil, M. L., Roach, C., Zhou, N., Morrisey, J. M., Vaidya, A. B., Bergman, L. W., and Hol, W. G. J. (2006). Structure of the MTIP-MyoA complex, a key component of the malaria parasite invasion motor. *Proceedings of the National Academy of Sciences of the United States of America*, 103(13):4852–7.

Bosch, J., Turley, S., Roach, C. M., Daly, T. M., Bergman, L. W., and Hol, W. G. J. (2007). The closed MTIP-myosin A-tail complex from the malaria parasite invasion machinery. *Journal of molecular biology*, 372(1):77–88.

Brandes, U., Delling, D., Gaertler, M., Gorke, R., Hoefer, M., Nikoloski, Z., and Wagner, D. (2008). On Modularity Clustering. *IEEE Transactions on Knowledge and Data Engineering*, 20(2):172–188.

Brinda, K. V., Kannan, N., and Vishveshwara, S. (2002). Analysis of homodimeric protein interfaces by graph-spectral methods. *Protein engineering*, 15(4):265–77.

Brinda, K. V. and Vishveshwara, S. (2005). A network representation of protein structures: implications for protein stability. *Biophysical journal*, 89(6):4159–70.

Brooks, B. and Karplus, M. (1983). Harmonic dynamics of proteins: Normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proceedings of the National Academy of Sciences of the United States of America*, 80(21):6571–6575.

Brooks, C. L., Karplus, M., and Pettitt, B. M. (1988). Proteins: A Theoretical Perspective of Dynamics, Structure, and Thermodynamics. *Advances in Chemical Physics*, 71:1–259.

Burge, S., Parkinson, G. N., Hazel, P., Todd, A. K., and Neidle, S. (2006). Quadruplex DNA: sequence, topology and structure. *Nucleic acids research*, 34(19):5402–15.

Caillet, J. and Claverie, P. (1975). Theoretical evaluations of the intermolecular interaction energy of a crystal: application to the analysis of crystal geometry. *Acta Crystallographica Section A*, 31(4):448–461.

Ceru, S., Sket, P., Prislan, I., Lah, J., and Plavec, J. (2014). A new pathway of DNA G-quadruplex formation. *Angewandte Chemie (International ed. in English)*, 53(19):4881–4.

Chaires, J. B. (2010). Human telomeric G-quadruplex: thermodynamic and kinetic studies of telomeric quadruplex stability. *The FEBS journal*, 277(5):1098–106.

Chennubhotla, C. and Bahar, I. (2006). Markov propagation of allosteric effects in biomolecular systems: application to GroEL-GroES. *Molecular systems biology*, 2:36.

Chennubhotla, C. and Bahar, I. (2007). Signal propagation in proteins and relation to equilibrium fluctuations. *PLoS computational biology*, 3(9):1716–26.

Chennubhotla, C., Rader, a. J., Yang, L.-W., and Bahar, I. (2005). Elastic network models for understanding biomolecular machinery: from enzymes to supramolecular assemblies. *Physical biology*, 2(4):S173–80.

Christen, M., Hünenberger, P. H., Bakowies, D., Baron, R., Bürgi, R., Geerke, D. P., Heinz, T. N., Kastenholz, M. a., Kräutler, V., Oostenbrink, C., Peter, C., Trzesniak, D., and van Gunsteren, W. F. (2005). The GROMOS software for biomolecular simulation: GROMOS05. *Journal of computational chemistry*, 26(16):1719–51.

Clackson, T. and Wells, J. A. (1995). A hot spot of binding energy in a hormone-receptor interface. *Science*, 267(5196):383–386.

Cleland, W. W., Andrews, T. J., Gutteridge, S., Hartman, F. C., and Lorimer, G. H. (1998). Mechanism of Rubisco: The Carbamate as General Base. *Chemical Reviews*, 98(2):549–562.

Cockrell, G. M. and Kantrowitz, E. R. (2012). Metal ion involvement in the allosteric mechanism of Escherichia coli aspartate transcarbamoylase. *Biochemistry*, 51(36):7128–37.

Cockrell, G. M., Zheng, Y., Guo, W., Peterson, A. W., Truong, J. K., and Kantrowitz, E. R. (2013). New paradigm for allosteric regulation of Escherichia coli aspartate transcarbamoylase. *Biochemistry*, 52(45):8036–47.

Collie, G. W. and Parkinson, G. N. (2011). The application of DNA and RNA G-quadruplexes to therapeutic medicines. *Chemical Society reviews*, 40(12):5867–92.

Costa, J. (2008). *Infinitesimal and combinatorial rigidity approaches to coarse grain proteins*. PhD thesis, Imperial College London.

Costa, J. R. and Yaliraki, S. N. (2006). Role of rigidity on the activity of proteinase inhibitors and their peptide mimics. *The journal of physical chemistry. B*, 110(38):18981–8.

Csermely, P., Korcsmáros, T., Kiss, H. J. M., London, G., and Nussinov, R. (2013). Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacology & therapeutics*, 138(3):333–408.

Csermely, P., Sandhu, K. S., Hazai, E., Hoksza, Z., Kiss, H. J. M., Miozzo, F., Veres, D. V., Piazza, F., and Nussinov, R. (2012). Disordered Proteins and Network Disorder in Network Descriptions of Protein Structure , Dynamics and Function : Hypotheses and a Comprehen- sive Review. *Current Protein and Peptide Science*, 13(1):19–33.

Dahiyat, B. I., Gordon, D. B., and Mayo, S. L. (1997). Automated design of the surface positions of protein helices. *Protein science : a publication of the Protein Society*, 6(6):1333–7.

Dai, J., Carver, M., and Yang, D. (2008). Polymorphism of human telomeric quadruplex structures. *Biochimie*, 90(8):1172–83.

De Cian, A., Lacroix, L., Douarre, C., Temime-Smaali, N., Trentesaux, C., Riou, J.-F., and Mergny, J.-L. (2008). Targeting telomeres and telomerase. *Biochimie*, 90(1):131–55.

De Los Rios, P. and Caldarelli, G. (2000). Putting proteins back into water. *Physical Review E*, 62(6):8449–8452.

del Sol, A., Araúzo-Bravo, M. J., Amoros, D., and Nussinov, R. (2007). Modular architecture of protein structures and allosteric communications: potential implications for signaling proteins and regulatory linkages. *Genome biology*, 8(5):R92.

del Sol, A., Fujihashi, H., Amoros, D., and Nussinov, R. (2006). Residues crucial for maintaining short paths in network communication mediate signaling in proteins. *Molecular systems biology*, 2:2006.0019.

del Sol, A. and O'Meara, P. (2005). Small-world network approach to identify key residues in protein-protein interaction. *Proteins*, 58(3):672–82.

DeLano, W. L. (2002). Unraveling hot spots in binding interfaces: progress and challenges. *Current opinion in structural biology*, 12(1):14–20.

Delmotte, A., Tate, E. W., Yaliraki, S. N., and Barahona, M. (2011). Protein multi-scale organization through graph partitioning and robustness analysis: application to the myosin-myosin light chain interaction. *Physical biology*, 8(5):055010.

Delvenne, J.-C., Schaub, M. T., Yaliraki, S. N., and Barahona, M. (2012). The stability of a graph partition: A dynamics-based framework for community detection. In Ganguly, N., Mukherjee, A., Choudhury, M., Peruani, F., and Mitra, B., editors, *Time Varying Dynamical Networks*. Birkhauser, Springer, To be published.

Delvenne, J.-C., Yaliraki, S. N., and Barahona, M. (2010). Stability of graph communities across time scales. SI. *Proceedings of the National Academy of Sciences of the United States of America*, 107(29):12755–12760.

Di Paola, L., De Ruvo, M., Paci, P., Santoni, D., and Giuliani, a. (2013). Protein contact networks: an emerging paradigm in chemistry. *Chemical reviews*, 113(3):1598–613.

Doruker, P., Atilgan, A. R., and Bahar, I. (2000). Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: application to alpha-amylase inhibitor. *Proteins: Structure, Function, and Genetics*, 40(3):512–24.

Douse, C. H., Green, J. L., Salgado, P. S., Simpson, P. J., Thomas, J. C., Langsley, G., Holder, A. a., Tate, E. W., and Cota, E. (2012). Regulation of the Plasmodium motor complex: phosphorylation of myosin A tail-interacting protein (MTIP) loosens its grip on MyoA. *The Journal of biological chemistry*, 287(44):36968–77.

Duff, A. P., Andrews, T. J., and Curmi, P. M. (2000). The transition between the open and closed states of rubisco is triggered by the inter-phosphate distance of the bound bisphosphate. *Journal of molecular biology*, 298(5):903–16.

Dumas, A. and Luedtke, N. W. (2010). Cation-mediated energy transfer in G-quadruplexes revealed by an internal fluorescent probe. *Journal of the American Chemical Society*, 132(51):18004–7.

Eaton, W. a., Henry, E. R., Hofrichter, J., Bettati, S., Viappiani, C., and Mozzarelli, A. (2007). Evolution of allosteric models for hemoglobin. *IUBMB life*, 59(8-9):586–99.

Eaton, W. A., Henry, E. R., Hofrichter, J., and Mozzarelli, A. (1999). Is cooperative oxygen binding by hemoglobin really understood? *Nature structural biology*, 6(4):351–8.

Eisenberg, E. and Hill, T. (1985). Muscle contraction and free energy transduction in biological systems. *Science*, 227(4690):999–1006.

Ellis, R. J. (2010). Biochemistry: Tackling unintelligent design. *Nature*, 463(7278):164–5.

Estrada, E. (2000). Characterization of 3D molecular structure. *Chemical Physics Letters*, 319(5-6):713–718.

Euler, L. (1736). Solutio problematis ad geometriam situs pertinentis. *Commentarii Academiae Scientiarum Imperialis Petropolitanae*, 8:128–140.

Farrow, R. E., Green, J., Katsimitsoulia, Z., Taylor, W. R., Holder, A. a., and Molloy, J. E. (2011). The mechanism of erythrocyte invasion by the malarial parasite, Plasmodium falciparum. *Seminars in cell & developmental biology*, 22(9):953–60.

Fernández-Recio, J. (2011). Prediction of protein binding sites and hot spots. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(5):680–698.

Field, C. B. (1998). Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components. *Science*, 281(5374):237–240.

Flory, P. J. (1969). *Statistical Mechanics of Chain Molecules*, volume 8. John Wiley & Sons, New York, USA.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3-5):75–174.

Frauenfelder, H., McMahon, B. H., Austin, R. H., Chu, K., and Groves, J. T. (2001). The role of structure, energy landscape, dynamics, and allostery in the enzymatic function of myoglobin. *Proceedings of the National Academy of Sciences of the United States of America*, 98(5):2370–4.

Frauenfelder, H., Sligar, S. G., and Wolynes, P. G. (1991). The energy landscapes and motions of proteins. *Science (New York, N.Y.)*, 254(5038):1598–1603.

Frénal, K., Polonais, V., Marq, J.-B., Stratmann, R., Limenitakis, J., and Soldati-Favre, D. (2010). Functional dissection of the apicomplexan glideosome molecular architecture. *Cell host & microbe*, 8(4):343–57.

Fuglebakk, E., Reuter, N., and Hinsen, K. (2013). Evaluation of Protein Elastic Network Models Based on an Analysis of Collective Motions. *Journal of Chemical Theory and Computation*, 9(12):5618–5628.

Gellert, M., Lipsett, M. N., and Davies, D. R. (1962). Helix formation by guanylic acid. *Proceedings of the National Academy of Sciences of the United States of America*, 48:2013–8.

Gerstein, M., Schulz, G., and Chothia, C. (1993). Domain closure in adenylate kinase. Joints on either side of two helices close like neighboring fingers. *Journal of molecular biology*, 229(2):494–501.

Gfeller, D. and De Los Rios, P. (2007). Spectral Coarse Graining of Complex Networks. *Physical Review Letters*, 99(3):038701.

Gilson, M. K. and Honig, B. H. (1986). The dielectric constant of a folded protein. *Biopolymers*, 25(11):2097–119.

Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12):7821—-26.

Go, N., Noguti, T., and Nishikawa, T. (1983). Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proceedings of the National Academy of Sciences of the United States of America*, 80(12):3696–3700.

Goldstein, H. (1953). *Classical mechanics*. Addison-Wesley, Cambridge.

Good, B. H., de Montjoye, Y.-A., and Clauset, A. (2010). Performance of modularity maximization in practical contexts. *Physical Review E*, 81(4):1–19.

Gray, R. D. and Chaires, J. B. (2008). Kinetics and mechanism of K+- and Na+-induced folding of models of human telomeric DNA into G-quadruplex structures. *Nucleic acids research*, 36(12):4191–4203.

Gray, R. D., Trent, J. O., and Chaires, J. B. (2014). Folding and unfolding pathways of the human telomeric G-quadruplex. *Journal of molecular biology*, 426(8):1629–50.

Green, J. L., Martin, S. R., Fielden, J., Ksagoni, A., Grainger, M., Yim Lim, B. Y. S., Molloy, J. E., and Holder, A. a. (2006). The MTIP-myosin A complex in blood stage malaria parasites. *Journal of molecular biology*, 355(5):933–41.

Greene, L. H. and Higman, V. a. (2003). Uncovering Network Systems Within Protein Structures. *Journal of Molecular Biology*, 334(4):781–791.

Gutteridge, S. (1991). The relative catalytic specificities of the large subunit core of Synechococcus ribulose bisphosphate carboxylase/oxygenase. *The Journal of biological chemistry*, 266(12):7359–7362.

Halle, B. (2002). Flexibility and packing in proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 99(3):1274–9.

Hamacher, K. and McCammon, J. A. (2006). Computing the Amino Acid Specificity of Fluctuations in Biomolecular Systems. *Journal of Chemical Theory and Computation*, 2(3):873–878.

Hartman, F. C. and Harpel, M. R. (1994). Structure, function, regulation, and assembly of D-ribulose-1,5-bisphosphate carboxylase/oxygenase. *Annual review of biochemistry*, 63:197–234.

Hayward, S. (2001). Normal Mode Analysis of Biological Molecules. In Becker, O. M. ., MacKerell Jr, A. D., Roux, B., and Watanabe, M., editors, *Computational Biochemistry and Biophysics*, chapter 8, pages 153–168. Marcel Dekker, Inc, New York, USA.

Hayward, S. and de Groot, B. L. (2008). Normal modes and essential dynamics. In Kukol, A., editor, *Methods in molecular biology, vol 443*, volume 443, pages 89–106. Humana Press, Totowa, NJ.

Heddi, B. and Phan, A. T. (2011). Structure of human telomeric DNA in crowded solution. *Journal of the American Chemical Society*, 133(25):9824–33.

Heintzelman, M. B. and Schwartzman, J. D. (1997). A novel class of unconventional myosins from Toxoplasma gondii. *Journal of molecular biology*, 271(1):139–46.

Henzler-Wildman, K. A. and Kern, D. (2007). Dynamic personalities of proteins. *Nature*, 450(7172):964–972.

Henzler-Wildman, K. A., Lei, M., Thai, V., Kerns, S. J., Karplus, M., and Kern, D. (2007a). A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature*, 450(7171):913–6.

Henzler-Wildman, K. a., Thai, V., Lei, M., Ott, M., Wolf-Watz, M., Fenn, T., Pozharski, E., Wilson, M. a., Petsko, G. a., Karplus, M., Hübner, C. G., and Kern, D. (2007b). Intrinsic motions along an enzymatic reaction trajectory. *Nature*, 450(7171):838–44.

Hess, B., Kutzner, C., van der Spoel, D., and Lindahl, E. (2008). GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *Journal of Chemical Theory and Computation*, 4(3):435–447.

Hettmann, C., Herm, a., Geiter, a., Frank, B., Schwarz, E., Soldati, T., and Soldati, D. (2000). A dibasic motif in the tail of a class XIV apicomplexan myosin is an essential determinant of plasma membrane localization. *Molecular biology of the cell*, 11(4):1385–400.

Hinsen, K. (1998). Analysis of domain motions by approximate normal mode calculations. *Proteins*, 33(3):417–29.

Huang, J. and Lipscomb, W. N. (2004). Aspartate transcarbamylase (ATCase) of Escherichia coli: a new crystalline R-state bound to PALA, or to product analogues citrate and phosphate. *Biochemistry*, 43(21):6415–21.

Hud, N. V. and Plavec, J. (2006). The Role of Cations in Determining Quadruplex Structure and Stability. In Neidle, S. and Balasubramanian, S., editors, *Quadruplex Nucleic Acids*, chapter 4, pages 100–130. The Royal Society of Chemistry, Cambridge.

Huheey, J. E., Keiter, E. A., and Keiter, R. L. (1993). *Inorganic Chemistry: Principles of Structure and Reactivity (4th Edition)*. HarperCollins College Publishers, New York, 4th edition.

Hunter, C. A. and Sanders, J. K. M. (1990). The nature of pi-pi interactions. *Journal of the American Chemical Society*, 112(14):5525–5534.

Ichiye, T. and Karplus, M. (1991). Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins*, 11(3):205–17.

Jacobs, D. and Thorpe, M. (1995). Generic Rigidity Percolation: The Pebble Game. *Physical Review Letters*, 75(22):4051–4054.

Jacobs, D. J. (1998). Generic rigidity in three-dimensional bond-bending networks. *Journal of Physics A: Mathematical and General*, 31(31):6653–6668.

Jacobs, D. J., Rader, A. J., Kuhn, L. A., and Thorpe, M. F. (2001). Protein flexibility predictions using graph theory. *Proteins*, 44(2):150–165.

James, H. M. (1947). Statistical Properties of Networks of Flexible Chains. *The Journal of Chemical Physics*, 15(9):651.

James, H. M. and Guth, E. (1943). Theory of the Elastic Properties of Rubber. *The Journal of Chemical Physics*, 11(10):455.

Jiao, X., Chang, S., Li, C.-h., Chen, W.-z., and Wang, C.-x. (2007). Construction and application of the weighted amino acid network based on energy. *Physical Review E*, 75(5):051903.

Jin, L., Stec, B., Lipscomb, W. N., and Kantrowitz, E. R. (1999). Insights into the mechanisms of catalysis and heterotropic regulation of Escherichia coli aspartate transcarbamoylase based upon a structure of the enzyme complexed with the bisubstrate analogue N-phosphonacetyl-L-aspartate at 2.1 A. *Proteins*, 37(4):729–42.

Jordan, D. B. and Ogren, W. L. (1981). Species variation in the specificity of ribulose biphosphate carboxylase/oxygenase. *Nature*, 291(5815):513–515.

Jordan, D. B. and Ogren, W. L. (1984). The CO2/O2 specificity of ribulose 1,5-bisphosphate carboxylase/oxygenase. *Planta*, 161(4):308–313.

Jorgensen, W. L. and Tirado-Rives, J. (1988). 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.

Kannan, N. and Vishveshwara, S. (1999). Identification of side-chain clusters in protein structures by a graph spectral method. *Journal of molecular biology*, 292(2):441–64.

Kantrowitz, E. R. (2012). Allostery and cooperativity in Escherichia coli aspartate transcarbamoylase. *Archives of biochemistry and biophysics*, 519(2):81–90.

Kapralov, M. V., Kubien, D. S., Andersson, I., and Filatov, D. a. (2011). Changes in Rubisco kinetics during the evolution of C4 photosynthesis in Flaveria (Asteraceae) are associated with positive selection on genes encoding the enzyme. *Molecular biology and evolution*, 28(4):1491–503.

Karkehabadi, S., Peddi, S. R., Anwaruzzaman, M., Taylor, T. C., Cederlund, A., Genkov, T., Andersson, I., and Spreitzer, R. J. (2005). Chimeric small subunits influence catalysis without causing global conformational changes in the crystal structure of ribulose-1,5-bisphosphate carboxylase/oxygenase. *Biochemistry*, 44(29):9851–61.

Karkehabadi, S., Taylor, T. C., and Andersson, I. (2003). Calcium Supports Loop Closure but not Catalysis in Rubisco. *Journal of Molecular Biology*, 334(1):65–73.

Karrer, B., Levina, E., and Newman, M. E. J. (2008). Robustness of community structure in networks. *Physical Review E*, 77(4):1–9.

Ke, H. M., Lipscomb, W. N., Cho, Y., and Honzatko, R. B. (1988). Complex of N-phosphonacetyl-L-aspartate with aspartate carbamoyltransferase. X-ray refinement, analysis of conformational changes and catalytic and allosteric mechanisms. *Journal of molecular biology*, 204(3):725–47.

Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H., and Phillips, D. C. (1958). A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. *Nature*, 181(4610):662–666.

Kim, K. H., Pan, Z., Honzatko, R. B., Ke, H.-m., and Lipscomb, W. N. (1987). Structural asymmetry in the CTP-liganded form of aspartate carbamoyltransferase from Escherichia coli. *Journal of Molecular Biology*, 196(4):853–875.

Klein, D. J. and Randić, M. (1993). Resistance distance. *Journal of Mathematical Chemistry*, 12(1):81–95.

Kloczkowski, A., Mark, J. E., and Erman, B. (1989). Chain dimensions and fluctuations in random elastomeric networks. 1. Phantom Gaussian networks in the undeformed state. *Macromolecules*, 22(3):1423–1432.

Knight, S., Andersson, I., and Brändén, C. I. (1990). Crystallographic analysis of ribulose 1,5-bisphosphate carboxylase from spinach at 2.4 A resolution. Subunit interactions and active site. *Journal of molecular biology*, 215(1):113–60.

Kortagere, S., Welsh, W. J., Morrisey, J. M., Daly, T., Ejigiri, I., Sinnis, P., Vaidya, A. B., and Bergman, L. W. (2010). Structure-based design of novel small-molecule inhibitors of Plasmodium falciparum. *Journal of chemical information and modeling*, 50(5):840–9.

Kosman, R. P., Gouaux, J. E., and Lipscomb, W. N. (1993). Crystal structure of CTP-ligated T state aspartate transcarbamoylase at 2.5 A resolution: implications for AT-Case mutants and the mechanism of negative cooperativity. *Proteins*, 15(2):147–76.

Kundu, S., Sorensen, D. C., and Phillips, G. N. (2004). Automatic domain decomposition of proteins by a Gaussian Network Model. *Proteins*, 57(4):725–33.

Laman, G. (1970). On graphs and rigidity of plane skeletal structures. *Journal of Engineering Mathematics*, 4(4):331–340.

Lambiotte, R. (2010). Multi-scale Modularity in Complex Networks. In *Proceedings of the 8th International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt)*, Avignon (France). IEEE.

Lambiotte, R., Delvenne, J.-C., and Barahona, M. (2009). Laplacian dynamics and multiscale modular structure in networks. *arXiv*, pages 1–29.

Lane, A. N., Chaires, J. B., Gray, R. D., and Trent, J. O. (2008). Stability and kinetics of G-quadruplex structures. *Nucleic acids research*, 36(17):5482–515.

Lee, J. Y., Okumus, B., Kim, D. S., and Ha, T. (2005). Extreme conformational diversity in human telomeric DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 102(52):18938–43.

Leitner, D. M., Havenith, M., and Gruebele, M. (2006). Biomolecule large-amplitude motion and solvation dynamics: modelling and probes from THz to X-rays. *International Reviews in Physical Chemistry*, 25(4):553–582.

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

Levy, Y. and Onuchic, J. N. (2006). Water mediation in protein folding and molecular recognition. *Annual review of biophysics and biomolecular structure*, 35:389–415.

Li, W., Hou, X.-M., Wang, P.-Y., Xi, X.-G., and Li, M. (2013). Direct measurement of sequential folding pathway and energy landscape of human telomeric G-quadruplex structures. *Journal of the American Chemical Society*, 135(17):6423–6.

Lin, M. S., Fawzi, N. L., and Head-Gordon, T. (2007). Hydrophobic Potential of Mean Force as a Solvation Function for Protein Structure Prediction. *Structure*, 15(6):727–740.

Lipscomb, W. N. and Kantrowitz, E. R. (2012). Structure and mechanisms of Escherichia coli aspartate transcarbamoylase. *Accounts of chemical research*, 45(3):444–53.

Liu, T., Whitten, S. T., and Hilser, V. J. (2007). Functional residues serve a dominant role in mediating the cooperativity of the protein ensemble. *Proceedings of the National Academy of Sciences of the United States of America*, 104(11):4347–52.

Lord Rayleigh (1919). On the problem of random vibrations, and of random flights in one, two, or three dimensions. *The London, Edinburgh and Dublin Phylosophical Magazine and Journal of Science*, 37(6):321.

Lorimer, G. H. and Andrews, T. J. (1973). Plant Photorespiration—An Inevitable Consequence of the Existence of Atmospheric Oxygen. *Nature*, 243(5406):359–360.

Lou, H. and Cukier, R. I. (2006). Molecular dynamics of apo-adenylate kinase: a principal component analysis. *The journal of physical chemistry. B*, 110(25):12796–808.

Lowey, S. and Trybus, K. M. (2010). Common structural motifs for the regulation of divergent class II myosins. *The Journal of biological chemistry*, 285(22):16403–7.

Lu, H.-M. and Liang, J. (2009). Perturbation-based Markovian transmission model for probing allosteric dynamics of large macromolecular assembling: a study of GroEL-GroES. *PLoS computational biology*, 5(10):e1000526.

Lyman, E., Pfaendtner, J., and Voth, G. a. (2008). Systematic multiscale parameterization of heterogeneous elastic network models of proteins. *Biophysical journal*, 95(9):4183–92.

Ma, B., Wolfson, H. J., and Nussinov, R. (2001). Protein functional epitopes: hot spots, dynamics and combinatorial libraries. *Current opinion in structural biology*, 11(3):364–369.

Ma, J. (2004). New Advances in Normal Mode Analysis of Supermolecular Complexes and Applications to Structural Refinement. *Current Protein and Peptide Science*, 5(2):119–123.

Ma, J. (2005). Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. *Structure*, 13(3):373–80.

Macol, C. P., Tsuruta, H., Stec, B., and Kantrowitz, E. R. (2001). Direct structural evidence for a concerted allosteric transition in Escherichia coli aspartate transcarbamoylase. *Nature structural biology*, 8(5):423–6.

Maizels, N. (2006). Dynamic roles for G4 DNA in the biology of eukaryotic cells. *Nature structural & molecular biology*, 13(12):1055–9.

Manning, G. S. (1978). The molecular theory of polyelectrolyte solutions with applications to the electrostatic properties of polynucleotides. *Quarterly Reviews of Biophysics*, 11(02):179.

Maragakis, P. and Karplus, M. (2005). Large amplitude conformational change in proteins explored with a plastic network model: adenylate kinase. *Journal of molecular biology*, 352(4):807–22.

Marchand, A., Ferreira, R., Tateishi-Karimata, H., Miyoshi, D., Sugimoto, N., and Gabelica, V. (2013). Sequence and solvent effects on telomeric DNA bimolecular G-quadruplex folding kinetics. *The journal of physical chemistry. B*, 117(41):12391–401.

Mashimo, T., Yagi, H., Sannohe, Y., Rajendran, A., and Sugiyama, H. (2010). Folding pathways of human telomeric type-1 and type-2 G-quadruplex structures. *Journal of the American Chemical Society*, 132(42):14910–14918.

Maslov, S. and Sneppen, K. (2002). Specificity and stability in topology of protein networks. *Science (New York, N.Y.)*, 296(5569):910–913.

Meila, M. (2003). Comparing Clusterings by the Variation of Information. In Schölkopf, B. and Warmuth, M., editors, *Learning Theory and Kernel Machines SE - 14*, volume 2777 of *Lecture Notes in Computer Science*, pages 173–187. Springer Berlin Heidelberg.

Meila, M. (2005). Comparing Clusterings âĂŞ An Axiomatic View. In *Proceedings of the 22nd international conference on Machine learning - ICML '05*, pages 577–584, Bonn, Germany, 2005. ACM Press.

Meila, M. (2007). Comparing clusteringsâĂŤan information based distance. *Journal of Multivariate Analysis*, 98(5):873–895.

Meliga, S. (2009). *Graph clustering of atomic networks for protein dynamics*. Master of research thesis, Imperial College London.

Mendes, K. R., Martinez, J. A., and Kantrowitz, E. R. (2010). Asymmetric allosteric signaling in aspartate transcarbamoylase. *ACS chemical biology*, 5(5):499–506.

Micheletti, C., Carloni, P., and Maritan, A. (2004). Accurate and efficient description of protein vibrational dynamics: comparing molecular dynamics and Gaussian models. *Proteins*, 55(3):635–45.

Ming, D. and Wall, M. E. (2006). Interactions in native binding sites cause a large change in protein dynamics. *Journal of molecular biology*, 358(1):213–23.

Molloy, M. and Reed, B. (1995). A critical point for random graphs with a given degree sequence. *Random Structures & Algorithms*, 6(2-3):161–180.

Monteith, J. L. and Moss, C. J. (1977). Climate and the Efficiency of Crop Production in Britain [and Discussion]. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 281(980):277–294.

Moreira, I. S., Fernandes, P. A., and Ramos, M. J. (2007). Hot spots–a review of the protein-protein interface determinant amino-acid residues. *Proteins*, 68(4):803–12.

Morrow, J. K. and Zhang, S. (2012). Computational Prediction of Protein Hot Spot Residues. *Current Pharmaceutical Design*, 18(9):1255–1265.

Moyzis, R. K., Buckingham, J. M., Cram, L. S., Dani, M., Deaven, L. L., Jones, M. D., Meyne, J., Ratliff, R. L., and Wu, J. R. (1988). A highly conserved repetitive DNA sequence, (TTAGGG)n, present at the telomeres of human chromosomes. *Proceedings of the National Academy of Sciences of the United States of America*, 85(18):6622–6.

Muirhead, H. and Perutz, M. F. (1963). Structure Of Hæmoglobin: A Three-Dimensional Fourier Synthesis of Reduced Human Haemoglobin at 5.5 âĎń Resolution. *Nature*, 199(4894):633–638.

Müller, C., Schlauderer, G., Reinstein, J., and Schulz, G. (1996). Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. *Structure*, 4(2):147–156.

Müller, C. W. and Schulz, G. E. (1992). Structure of the complex between adenylate kinase from Escherichia coli and the inhibitor Ap5A refined at 1.9 A resolution. A model for a catalytic transition state. *Journal of molecular biology*, 224(1):159–77.

Neidle, S. (2009). The structures of quadruplex nucleic acids and their drug complexes. *Current opinion in structural biology*, 19(3):239–250.

Neidle, S. (2010). Human telomeric G-quadruplex: the current status of telomeric G-quadruplexes as therapeutic targets in human cancer. *The FEBS journal*, 277(5):1118–25.

Neidle, S. and Balasubramanian, S., editors (2006). *Quadruplex Nucleic Acids*. Royal Society of Chemistry, Cambridge.

Newell, J. O., Markby, D. W., and Schachman, H. K. (1989). Cooperative binding of the bisubstrate analog N-(phosphonacetyl)-L-aspartate to aspartate transcarbamoylase and the heterotropic effects of ATP and CTP. *The Journal of biological chemistry*, 264(5):2476–81.

Newman, M. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69(2):26113.

Newman, M. E. J. (2005). Chapter 2. Random graphs as models of networks. In Bornholdt, S. and Schuster, H. G., editors, *Handbook of Graphs and Networks*, pages 35—-68. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, FRG.

Olsson, U. and Wolf-Watz, M. (2010). Overlap between folding and functional energy landscapes for adenylate kinase conformational change. *Nature communications*, 1(8):111.

Papoian, G. a., Ulander, J., Eastwood, M. P., Luthey-Schulten, Z., and Wolynes, P. G. (2004). Water in protein structure prediction. *Proceedings of the National Academy of Sciences of the United States of America*, 101(10):3352–7.

Park, K. and Kim, D. (2011). Modeling allosteric signal propagation using protein structure networks. *BMC bioinformatics*, 12 Suppl 1(Suppl 1):S23.

Parry, M. A. J., Keys, A. J., and Gutteridge, S. (1989). Variation in the Specificity Factor of C 3 Higher Plant Rubiscos Determined by the Total Consumption of Ribulose-P 2. *Journal of Experimental Botany*, 40(3):317–320.

Parry, M. A. J., Keys, A. J., Madgwick, P. J., Carmo-Silva, A. E., and Andralojc, P. J. (2008). Rubisco regulation: a role for inhibitors. *Journal of experimental botany*, 59(7):1569–80.

Parry, M. A. J., Madgwick, P. J., Carvalho, J. F. C., and Andralojc, P. J. (2007). Prospects for increasing photosynthesis by overcoming the limitations of Rubisco. *The Journal of Agricultural Science*, 145(01):31–43.

Perutz, M. F. (1970). Stereochemistry of Cooperative Effects in Haemoglobin: Haem-Haem Interaction and the Problem of Allostery. *Nature*, 228(5273):726–734.

Petri, G. and Expert, P. (2014). Temporal stability of network partitions. *arXiv [1404.7170]*, page 15.

Phan, A. T. (2010). Human telomeric G-quadruplex: structures of DNA and RNA sequences. *The FEBS journal*, 277(5):1107–17.

Phan, A. T., Kuryavyi, V., and Patel, D. J. (2006). DNA architecture: from G to Z. *Current opinion in structural biology*, 16(3):288–298.

Portis Jr, A. R. (1992). Regulation of Ribulose 1,5-Bisphosphate Carboxylase/Oxygenase Activity. *Annual Review of Plant Physiology and Plant Molecular Biology*, 43(1):415–437.

Rader, A. J., Chennubhotla, C., Yang, L.-W., and Bahar, I. (2006). *The Gaussian Network Model: theory and applications*, chapter 3, pages 41–64. Chapman & Hall/CRC.

Rader, A. J., Hespenheide, B. M., Kuhn, L. A., and Thorpe, M. F. (2002). Protein unfolding: rigidity lost. *Proceedings of the National Academy of Sciences of the United States of America*, 99(6):3540–5.

Radivojac, P., Iakoucheva, L. M., Oldfield, C. J., Obradovic, Z., Uversky, V. N., and Dunker, A. K. (2007). Intrinsic disorder and functional proteomics. *Biophysical journal*, 92(5):1439–56.

Rand, W. M. (1971). Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, 66(336):846–850.

Ravishanker, G., Auffinger, P., Langley, D. R., Jayaram, B., Young, M. A., and Beveridge, D. L. (2007). Treatment of Counterions in Computer Simulations of DNA. In Lipkowitz, K. B. and Boyd, D. B., editors, *Reviews in Computational Chemistry (Volume 11)*, chapter 6, pages 317–372. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Rayment, I., Holden, H., Whittaker, M., Yohn, C., Lorenz, M., Holmes, K., and Milligan, R. (1993). Structure of the actin-myosin complex and its implications for muscle contraction. *Science*, 261(5117):58–65.

Read, B. A. and Tabita, F. R. (1994). High substrate specificity factor ribulose bisphosphate carboxylase/oxygenase from eukaryotic marine algae and properties of recombinant cyanobacterial RubiSCO containing "algal" residue modifications. *Archives of biochemistry and biophysics*, 312(1):210–8.

Rees-Channer, R. R., Martin, S. R., Green, J. L., Bowyer, P. W., Grainger, M., Molloy, J. E., and Holder, A. a. (2006). Dual acylation of the 45 kDa gliding-associated protein (GAP45) in Plasmodium falciparum merozoites. *Molecular and biochemical parasitology*, 149(1):113–6.

Reuveni, S., Granek, R., and Klafter, J. (2010a). Anomalies in the vibrational dynamics of proteins are a consequence of fractal-like structure. *Proceedings of the National Academy of Sciences of the United States of America*, 107(31):13696–700.

Reuveni, S., Granek, R., and Klafter, J. (2010b). General mapping between random walks and thermal vibrations in elastic networks: Fractal networks as a case study. *Physical Review E*, 82(4):1–4.

Ribeiro, A. A. S. T. and Ortiz, V. (2014). Determination of Signaling Pathways in Proteins through Network Theory: Importance of the Topology. *Journal of Chemical Theory and Computation*, 10(4):1762–1769.

Ronhovde, P. and Nussinov, Z. (2009). Multiresolution community detection for megascale networks by information-based replica correlations. *Physical Review E*, 80(1):1–18.

Sage, R. F. (2002). Variation in the k(cat) of Rubisco in C(3) and C(4) plants and some implications for photosynthetic performance at high and low temperature. *Journal of experimental botany*, 53(369):609–20.

Schaeffer, S. E. (2007). Graph clustering. *Computer Science Review*, 1(1):27–64.

Schaub, M. T., Delvenne, J.-C., Yaliraki, S. N., and Barahona, M. (2012). Markov Dynamics as a Zooming Lens for Multiscale Community Detection: Non Clique-Like Communities and the Field-of-View Limit. *PloS one*, 7(2):e32210.

Schaub, M. T., Lehmann, J., Yaliraki, S. N., and Barahona, M. (2014). Structure of complex networks: Quantifying edge-to-edge relations by failure-induced flow redistribution. *Network Science*, pages 1–24.

Schiaffino, S. and Reggiani, C. (1996). Molecular diversity of myofibrillar proteins: gene regulation and functional significance. *Physiological reviews*, 76(2):371–423.

Schliwa, M. and Woehlke, G. (2003). Molecular motors. *Nature*, 422(6933):759–65.

Schneider, G., Lindqvist, Y., and Brändén, C. I. (1992). RUBISCO: structure and mechanism. *Annual review of biophysics and biomolecular structure*, 21:119–43.

Schrödinger, LLC (2010). The {PyMOL} Molecular Graphics System, Version˜1.3r1.

Schüttelkopf, A. W. and van Aalten, D. M. F. (2004). PRODRG: a tool for high-throughput crystallography of protein-ligand complexes. *Acta crystallographica. Section D, Biological crystallography*, 60(Pt 8):1355–63.

Sen, D. and Gilbert, W. (1988). Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis. *Nature*, 334(6180):364–6.

Sethi, A., Eargle, J., Black, A. a., and Luthey-Schulten, Z. (2009). Dynamical networks in tRNA:protein complexes. *Proceedings of the National Academy of Sciences of the United States of America*, 106(16):6620–5.

Shay, J. W. and Wright, W. E. (2011). Role of telomeres and telomerase in cancer. *Seminars in cancer biology*, 21(6):349–53.

Sinitskiy, A. V., Saunders, M. G., and Voth, G. a. (2012). Optimal number of coarse-grained sites in different components of large biomolecular complexes. *The journal of physical chemistry. B*, 116(29):8363–74.

Snow, R. W., Guerra, C. A., Noor, A. M., Myint, H. Y., and Hay, S. I. (2005). The global distribution of clinical episodes of Plasmodium falciparum malaria. *Nature*, 434(7030):214–7.

Soheilifard, R., Makarov, D. E., and Rodin, G. J. (2008). Critical evaluation of simple network models of protein dynamics and their comparison with crystallographic B-factors. *Physical biology*, 5(2):026008.

Sponer, J., Riley, K. E., and Hobza, P. (2008). Nature and magnitude of aromatic stacking of nucleic acid bases. *Physical chemistry chemical physics : PCCP*, 10(19):2595–610.

Spreitzer, R. (1999). Questions about the complexity of chloroplast ribulose-1,5-bisphosphate carboxylase/oxygenase. *Photosynthesis Research*, 60(1):29–42.

Spreitzer, R. (2003). Role of the small subunit in ribulose-1,5-bisphosphate carboxylase/oxygenase. *Archives of Biochemistry and Biophysics*, 414(2):141–149.

Spreitzer, R. J. and Salvucci, M. E. (2002). Rubisco: structure, regulatory interactions, and possibilities for a better enzyme. *Annual review of plant biology*, 53:449–75.

Stadlbauer, P., Krepl, M., Cheatham III, T. E., Koca, J., and Sponer, J. (2013). Structural dynamics of possible late-stage intermediates in folding of quadruplex DNA studied by molecular simulations. *Nucleic acids research*, 41(14):7128–43.

Stevens, R. C., Gouaux, J. E., and Lipscomb, W. N. (1990). Structural consequences of effector binding to the T state of aspartate carbamoyltransferase: crystal structures of the unligated and ATP- and CTP-complexed enzymes at 2.6-.ANG. resolution. *Biochemistry*, 29(33):7691–7701.

Stevens, R. C. and Lipscomb, W. N. (1992). A molecular mechanism for pyrimidine and purine nucleotide control of aspartate transcarbamoylase. *Proceedings of the National Academy of Sciences*, 89(12):5281–5285.

Stewart, S. a. and Weinberg, R. a. (2006). Telomeres: cancer to human aging. *Annual review of cell and developmental biology*, 22:531–57.

Strogatz, S. H. (2001). Exploring complex networks. *Nature*, 410(6825):268–76.

Sweeney, H. L. and Houdusse, A. (2010). Structural and functional insights into the Myosin motor mechanism. *Annual review of biophysics*, 39:539–57.

Swigon, D. (2009). The Mathematics of DNA Sturcture, Mechanics, and Dynamics. In Benham, C. J., Harvey, S., Olson, W. K., Sumners, D. W., and Swigon, D., editors, *Mathematics of DNA Structure, Function and Interactions*, volume 150 of *The IMA Volumes in Mathematics and its Applications*, pages 293–320. Springer New York.

Tabita, F. R. (1999). Microbial ribulose 1,5-bisphosphate carboxylase/oxygenase: A different perspective. *Photosynthesis Research*, 60(1):1–28.

Tabita, F. R., Satagopan, S., Hanson, T. E., Kreel, N. E., and Scott, S. S. (2008). Distinct form I, II, III, and IV Rubisco proteins from the three kingdoms of life provide clues about Rubisco evolution and structure/function relationships. *Journal of experimental botany*, 59(7):1515–24.

Tama, F. and Sanejouand, Y. H. (2001). Conformational change of proteins arising from normal mode calculations. *Protein engineering*, 14(1):1–6.

Tanner, J. J., Smith, P. E., and Krause, K. L. (1993). Molecular dynamics simulations and rigid body (TLS) analysis of aspartate carbamoyltransferase: evidence for an uncoupled R state. *Protein science : a publication of the Protein Society*, 2(6):927–35.

Taylor, T. C. and Andersson, I. (1996). Structural transitions during activation and ligand binding in hexadecameric Rubisco inferred from the crystal structure of the activated unliganded spinach enzyme. *Nature Structural Biology*, 3(1):95–101.

Taylor, T. C. and Andersson, I. (1997a). Structure of a product complex of spinach ribulose-1,5-bisphosphate carboxylase/oxygenase. *Biochemistry*, 36(13):4041–6.

Taylor, T. C. and Andersson, I. (1997b). The structure of the complex between rubisco and its natural substrate ribulose 1,5-bisphosphate. *Journal of molecular biology*, 265(4):432–44.

Taylor, T. C., Fothergill, M. D., and Andersson, I. (1996). A common structural basis for the inhibition of ribulose 1,5-bisphosphate carboxylase by 4-carboxyarabinitol 1,5-bisphosphate and xylulose 1,5-bisphosphate. *The Journal of biological chemistry*, 271(51):32894–9.

Tcherkez, G. G. B., Farquhar, G. D., and Andrews, T. J. (2006). Despite slow catalysis and confused substrate specificity, all ribulose bisphosphate carboxylases may be nearly perfectly optimized. *Proceedings of the National Academy of Sciences of the United States of America*, 103(19):7246–51.

Thomas, J. C., Green, J. L., Howson, R. I., Simpson, P., Moss, D. K., Martin, S. R., Holder, A. a., Cota, E., and Tate, E. W. (2010). Interaction and dynamics of the Plasmodium falciparum MTIP-MyoA complex, a key component of the invasion motor in the malaria parasite. *Molecular bioSystems*, 6(3):494–8.

Thorpe, M. (2009). FIRST 6.2.1 user guide. Technical report.

Thorpe, M. F. (2007). Comment on elastic network models and proteins. *Physical biology*, 4(1):60–3; discussion 64–5.

Timson, D. (2003). Fine tuning the myosin motor: the role of the essential light chain in striated muscle myosin. *Biochimie*, 85(7):639–645.

Tirion, M. M. (1996). Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Physical Review Letters*, 77(9):1905–1908.

Tozzini, V. (2005). Coarse-grained models for proteins. *Current opinion in structural biology*, 15(2):144–50.

Tran, P. L. T., De Cian, A., Gros, J., Moriyama, R., and Mergny, J.-L. (2013). Tetramolecular quadruplex stability and assembly. *Topics in current chemistry*, 330:243–73.

Trybus, K. M. (1994). Role of myosin light chains. *Journal of muscle research and cell motility*, 15(6):587–94.

Turley, S., Khamrui, S., Bergman, L. W., and Hol, W. G. J. (2013). The compact conformation of the Plasmodium knowlesi myosin tail interacting protein MTIP in complex with the C-terminal helix of myosin A. *Molecular and biochemical parasitology*, 190(2):56–9.

van Dongen, S. (2000). A cluster algorithm for graphs. *Information Systems [INS]*, (R 0010):1–40.

van Lun, M., van der Spoel, D., and Andersson, I. (2011). Subunit interface dynamics in hexadecameric rubisco. *Journal of molecular biology*, 411(5):1083–98.

Vendruscolo, M., Dokholyan, N., Paci, E., and Karplus, M. (2002). Small-world view of the amino acids that play a key role in protein folding. *Physical Review E*, 65(6):061910.

Vijayabaskar, M. S. and Vishveshwara, S. (2010). Interaction energy based protein structure networks. *Biophysical journal*, 99(11):3704–15.

Vishveshwara, S., Brinda, K. V., and Kannan, N. (2002). Protein Structure: Insights From Graph Theory. *Journal of Theoretical and Computational Chemistry*, 01(01):187–211.

Wang, J., Stieglitz, K. a., Cardia, J. P., and Kantrowitz, E. R. (2005). Structural basis for ordered substrate binding and cooperativity in aspartate transcarbamoylase. *Proceedings of the National Academy of Sciences of the United States of America*, 102(25):8881–6.

Warshel, A., Sharma, P. K., Kato, M., and Parson, W. W. (2006). Modeling electrostatic effects in proteins. *Biochimica et biophysica acta*, 1764(11):1647–76.

Wells, J. A. (1991). Systematic mutational analyses of protein-protein interfaces. *Methods in enzymology*, 202:390–411.

Whitney, S. M., Houtz, R. L., and Alonso, H. (2011). Advancing our understanding and capacity to engineer nature's CO2-sequestering enzyme, Rubisco. *Plant physiology*, 155(1):27–35.

Williamson, J. R., Raghuraman, M. K., and Cech, T. R. (1989). Monovalent cation-induced structure of telomeric DNA: the G-quartet model. *Cell*, 59(5):871–80.

Wilson, E. B., Decius, J. C., and Cross, P. C. (1955). *Molecular vibrations: the theory of infrared and Raman vibrational spectra*. McGraw-Hill Book Company, York, Pa, USA.

Winlund, C. C. and J. Chamberlin, M. (1970). Binding of cytidine triphosphate to aspartate transcarbamylase. *Biochemical and Biophysical Research Communications*, 40(1):43–49.

Word, J. M., Lovell, S. C., Richardson, J. S., and Richardson, D. C. (1999). Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *Journal of molecular biology*, 285(4):1735–47.

Wu, F. and Huberman, B. a. (2004). Finding communities in linear time: a physics approach. *The European Physical Journal B - Condensed Matter*, 38(2):331–338.

Wyatt, J. R., Davis, P. W., and Freier, S. M. (1996). Kinetics of G-quartet-mediated tetramer formation. *Biochemistry*, 35(24):8002–8.

Xu, Y., Xu, D., and Gabow, H. N. (2000). Protein domain decomposition using a graph-theoretic approach. *Bioinformatics*, 16(12):1091–1104.

Xue, Y., Liu, J.-q., Zheng, K.-w., Kan, Z.-y., Hao, Y.-h., and Tan, Z. (2011). Kinetic and thermodynamic control of G-quadruplex folding. *Angewandte Chemie (International ed. in English)*, 50(35):8046–50.

Yaliraki, S. N. and Barahona, M. (2007). Chemistry across scales: from molecules to cells. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 365(1861):2921–34.

Yang, L.-W. and Bahar, I. (2005). Coupling between catalytic site and collective dynamics: a requirement for mechanochemical activity of enzymes. *Structure (London, England : 1993)*, 13(6):893–904.

Yang, L.-W., Eyal, E., Chennubhotla, C., Jee, J., Gronenborn, A. M., and Bahar, I. (2007). Insights into equilibrium dynamics of proteins from comparison of NMR and X-ray data with computational predictions. *Structure (London, England : 1993)*, 15(6):741–9.

Yokota, a., Higashioka, M., and Wadano, a. (1991). Cooperative binding of carboxyarabinitol bisphosphate to the regulatory sites of ribulose bisphosphate carboxylase/oxygenase from spinach. *Journal of biochemistry*, 110(2):253–256.

Yu, H., Gu, X., Nakano, S.-i., Miyoshi, D., and Sugimoto, N. (2012). Beads-on-a-string structure of long telomeric DNAs under molecular crowding conditions. *Journal of the American Chemical Society*, 134(49):20060–9.

Zelitch, I. (1973). Plant Productivity and the Control of Photorespiration. *Proceedings of the National Academy of Sciences of the United States of America*, 70(2):579–584.

Zhang, A. Y. Q. and Balasubramanian, S. (2012). The kinetics and folding pathways of intramolecular G-quadruplex nucleic acids. *Journal of the American Chemical Society*, 134(46):19297–308.

Zhang, Z., Dai, J., Veliath, E., Jones, R. A., and Yang, D. (2010). Structure of a two-G-tetrad intramolecular G-quadruplex formed by a variant human telomeric sequence in K+ solution: insights into the interconversion of human telomeric G-quadruplex structures. *Nucleic acids research*, 38(3):1009–21.

Zhu, H., Xiao, S., and Liang, H. (2013). Structural dynamics of human telomeric G-quadruplex loops studied by molecular dynamics simulations. *PloS one*, 8(8):e71380.

Zhu, X.-G., Portis Jr, A. R., and Long, S. P. (2004). Would transformation of C3 crop plants with foreign Rubisco increase productivity? A computational analysis extrapolating from kinetic properties to canopy photosynthesis. *Plant, Cell and Environment*, 27(2):155–165.